

The replicability of spatially-resolved transcriptomics for modern neuroscience

SHAINA LU

*A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy*

School of Biological Sciences
Cold Spring Harbor Laboratory

September 23, 2021

“What we play is life.”

Louis Armstrong

Acknowledgements

First and foremost, I would like to thank my thesis advisors Tony Zador and Jesse Gillis. Though I'm still far from considering myself a neuroscientist, thanks to Tony for taking me on without any knowledge of the topic. I have deeply benefited from his clarity of thought and infinite patience. To Jesse, I have infinite gratitude for taking a second chance on me a year later than I traditionally would have joined his lab. I am deeply indebted to his genuine care for his mentees and his always honest, though sometimes brutal, feedback.

I would also like to thank my thesis committee members: Molly Gale Hammell (chair), Jessica Tollkuhn, Leemor Joshua-Tor (academic mentor), and formerly Adam Kepecs. They have always provided excellent feedback and support, keeping my best interests at heart. To Leemor, I have deep gratitude for her formidable advising that always forced me to productively gather my thoughts and concerns to best navigate graduate school. I thank also my external examiner, Adam Hantman, for agreeing to take on the additional commitment of a thesis defense especially during the pandemic. Further, I thank the members of the School of Biological Sciences at Cold Spring Harbor Laboratory (CSHL): Alex Gann, Alyson Kass-Eisler, Monn Monn Myat, Kim Creteur, and Kim Graham. Their door is always open for guidance, logistics, advice, and, of course, candy. The lack of any of these would have yielded my PhD unteneable. To Alex, thanks for convincing me not to quit my PhD research on that fateful fall day.

To my labmates in both the Gillis and Zador labs, you are too many to name, but I am forever grateful for your camaraderie, scientific discussion, technical support, advice, and friendship. Thank you also for always providing a willing, and sometimes not, ear to listen to me complain about it all. I will always hold dearly the freewheeling lunch discussions, lab excursions, and everything else in between. A special thanks to those in my lab and beyond that I have been lucky enough to collaborate with on specific projects: Xiaoyin Chen, Stephan Fischer, Nathan Fox, Daniel Fürth, Dinos Meletis, Cantin Ortiz, Yu-Chi Sun, and Ching Zhan.

I also thank the Women in Science and Engineering (WiSE) group and all those

involved that I had the pleasure of working with and befriending through a shared passion project. Thanks also to all my classmates and other CSHL friends for providing many fond memories, friendly faces, and perspective throughout. Thank you to my friend and housemate Lyndsey Aguirre for providing me with endless baked goods and gracefully navigating the early scares of COVID-19 together. I would also like to thank my lifelong friends scattered around the globe for providing me a much needed support system, escape outside of graduate school, and perspective from the 'real world'. I truly admire all you all have accomplished as young adults and consider you all as close confidants for my personal and professional life.

Thanks is not enough for my immediate and extended family, across both oceans, for the endless support and cheer. To my parents and grandparents, thank you for always supporting, encouraging, and loving me even when I was truly a brat. Thanks also to my parents for always prioritizing my education through natural disasters and health crises. To my sister, thanks for always being my biggest role model and stepping in where our parents could not. Thank you to Diogo Maia e Silva for your endless support and love, though sometimes tough. In addition to always giving me a reason to smile and stay sane, to my dog Kai, though you did not understand a word, thanks for listening to me practice every single talk of my graduate school career. Your ears were always telling of when I had droned on for too long. (Too bad he can not also proof read this!)

Contents

Acknowledgements	ii
List of Figures	viii
List of Tables	xi
List of Abbreviations	xii
1 Introduction I: Biology in the era of big data	1
1.1 Neuroscience in the era of big biology	1
1.2 Biology in the era of big biology	2
1.2.1 Introduction to big biology research	2
1.2.2 Diversity of consortia	5
1.3 Consortia have enabled massive data collection and tool development	7
1.3.1 Data	7
1.3.2 Data sharing	8
1.3.3 Democratizing new technologies and computational software	10
1.4 Robustness of new technologies is critical	12
1.5 Conclusions	13
1.6 Thesis Outline	14
2 Introduction II: Spatially resolved transcriptomics	16
2.1 Spatially-resolved transcriptomics is poised to be transformative across bi- ology	16
2.2 Spatial techniques, experimental	17

2.2.1	Capture-based methods	18
2.2.2	<i>In-situ</i> hybridization	25
2.2.3	<i>In situ</i> sequencing	27
2.2.4	Microdissection and other methods	29
2.2.5	Trade-offs of the different spatially resolved transcriptomics approaches and concluding thoughts	31
2.3	Spatial techniques, computational	33
2.3.1	Pre-processing of spatial data	34
2.3.2	Assay-independent spatial analysis frameworks	35
2.3.3	Identification of spatially differentially expressed or spatial marker genes	37
2.3.4	Integration of spatial data with single-cell RNA-seq	39
2.3.5	Deconvolution of non-single-cell resolution spatial data	41
2.3.6	Mapping spatial information onto single-cell data	42
2.3.7	Conclusions	43
2.4	Spatial expression combined with other data modalities in neuroscience	45
2.4.1	Early spatial expression in neuroscience (Allen Brain Atlas)	46
2.4.2	A few examples of multi-modal neuroscience studies	47
2.4.3	New experimental tools for multi-modal neuroscience	49
2.5	The need for benchmarking across spatial techniques	50
3	Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain	52
3.1	Introduction	52
3.2	Results and discussion	55
3.2.1	Allen Reference Atlas brain areas are classifiable using gene expression alone	55
3.2.2	Cross-dataset learning of Allen Reference Atlas brain areas	67
3.2.3	Distance in semantic space, but not physical space provides a potential explanation for cross-dataset performance	76

3.2.4	Finding a uniquely identifying gene expression profile for individual brain areas	81
3.3	Methods	85
3.4	Conclusions	93
3.5	Future directions	96
4	Vignettes of spatial transcriptomics in neuroscience	98
4.1	Other approaches to spatially-resolved transcriptomics in neuroscience . . .	98
4.2	Replicability of BARseq2 with re-quantified ABA ISH	99
4.2.1	Brief introduction to cadherins	99
4.2.2	Brief introduction to the auditory cortex	99
4.2.3	Re-quantification of cadherins from raw in situ hybridization images	101
4.2.4	Methods	102
4.3	Probing the relationship between spatial and single cell data in the primary motor cortex	105
4.3.1	Brief introduction to cell-type markers defined from scRNA-seq . .	105
4.3.2	Brief introduction to the motor cortex	105
4.3.3	In the primary motor cortex, are spatial expression patterns merely capturing cell type composition differences across the brain?	106
4.3.4	Methods	108
4.4	Reading out sequences for <i>in situ</i> sequencing	109
4.5	Discussion	115
5	Conclusions and perspectives	117
5.1	The potential of spatially-resolved transcriptomics	117
5.2	Summary and conclusions	118
5.3	Future directions and broader impact	119
5.4	Biology and big data are one in the same	120
A	Unpublished Review: Consortia	122

B Nature Methods News & Views: Integrative analysis methods to bridge trade-offs in spatial transcriptomics data	134
C PLOS Biology: Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain	137
D Nature Neuroscience Technical Report: Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections	171
Bibliography	201

List of Figures

1.1	Lifespan of a selected group of consortia.	3
2.1	Simplified schematic of most capture-based spatially-resolved transcriptomics methods.	19
2.2	Simplified schematic of microdissection and other manual tissue selection spatially-resolved transcriptomics methods.	30
2.3	Schematic of SpaGCN spatial analysis tool.	38
2.4	Schematic of Tangram spatial analysis tool.	41
3.1	Collection and processing of spatial gene expression datasets.	56
3.2	Canonical brain areas are classifiable using gene expression alone in the ABA and ST datasets.	59
3.3	Histogram visualization of LASSO ($\lambda = 0.1$) performance.	60
3.4	Additional visualization and verification of within dataset LASSO results with a new random train/test split.	61
3.5	LASSO performance with permuted labels falls to chance.	62
3.6	Classifying brain areas using k-NN.	63
3.7	Classification using single genes, relative expression across datasets, and PCA.	64
3.8	Internal data structure of ABA and ST datasets.	66
3.9	Sample correlation within brain areas and relationship between size and classification performance.	68
3.10	Cross-dataset learning shows that models do not generalize bi-directionally.	70

3.11	Additional verification of cross dataset LASSO results with a new random train/test split.	71
3.12	Visualization of the ABA and ST datasets together in low-dimensional space.	72
3.13	Feature set sizes for correlation-based feature selection (CFS) and cross-dataset results for CFS with averaging across feature sets.	74
3.14	Three-way cross-dataset LASSO shows that the sagittal subset of the ABA is the most distinct.	76
3.15	Summary plots for cross dataset analysis of ST, ABA coronal, and ABA sagittal with various parameterizations.	77
3.16	Spatial expression patterns reflect distance in semantic space, but not physical distance in the brain.	78
3.17	Comparison of path length and Euclidean distance to LASSO performance for various parameterizations of LASSO.	79
3.18	Identifying a unique gene expression profile for individual brain areas. . . .	82
3.19	Leaf brain area expression profiles are identifiable within dataset, but do not generalize cross dataset.	83
3.20	One v. all and one v. one analysis across various parameterizations.	84
3.21	Relationship between sample size and LASSO hyperparameter choice. . . .	90
4.1	Comparison of <i>Cdh8</i> , <i>Pcdh19</i> , and <i>Pcdh20</i> BARseq expression with RNAscope in primary auditory cortex.	100
4.2	Expression energy from Allen of <i>Cdh8</i> , <i>Pcdh19</i> , and <i>Pcdh20</i> across layers of the primary auditory cortex.	101
4.3	Comparison between BARseq2 and ABA <i>in situ</i> expression atlas for cadherins.	103
4.4	Summary of relative gene expression observed using BARseq2 and in Allen gene expression atlas.	104
4.5	The primary motor cortex is classifiable using gene expression.	106
4.6	Cross-dataset performance of identifying the primary motor cortex	107
4.7	Motor cortex gene markers are not enriched for broad cell types.	108

4.8	Motor cortex gene markers are not enriched for GABAergic and excitatory sub-types.	108
4.9	Base-calling schematic for FISSEQ.	110
4.10	Reference-free base-calling schematic for <i>in situ</i> sequencing.	110
4.11	Schematic illustrating magnitude-independence of cosine distance.	111
4.12	Trial-and-error testing of various thresholding approaches to identify pixels corresponding to sequenced reads.	113
4.13	Trial-and-error determination of threshold for base-calling.	113
4.14	Handling ties in maximum intensity for base-calling.	114
4.15	Pseudo-colored grouping of pixels sharing the same sequence.	114
4.16	Morphological erosion of clustered pixels.	114
4.17	Application of base-calling pipeline to an independent dataset.	115

List of Tables

2.1	Comparison of types of spatially resolved transcriptomics approaches. . . .	18
2.2	Comparison of spatial resolution given by spot density in capture-based spatially resolved transcriptomics techniques.	20
2.3	Comparison of capture sensitivity in capture-based spatially resolved transcriptomics techniques.	20
2.4	Examples of computational spatially resolved transcriptomics tools. . . .	34

List of Abbreviations

ABA	Allen B rain Atlas
ARA	Allen R eference Atlas
AUROC	a rea u nder the r eceiver o perating characteristic curve
BaristaSeq	b arcoded <i>in situ</i> targeted s equencing
BARseq	b arcoded a natomy r esolved by s equencing
BICCN	B RAIN Initiative Cell Census N etwork
bp	b ase p airs
cDNA	complementary D N
CFS	correlation-based f eature selection
DBiT-seq	d eterministic b arcoding i n tissue for spatial omics s equencing
DE	d ifferential e xpression or d ifferentially e xpressed
(DNA) seqFISH(+)	(D N) s equential f luorescence <i>in situ</i> h ybridization
FISSEQ	fluorescent <i>in situ</i> RNA s equencing
HCA	H uman C ell Atlas
HDST	H igh- D ensity S patial T ranscriptomics
HGP	H uman G enome P roject
ISH	<i>in situ</i> h ybridization
ISS	<i>in situ</i> sequencing
k-NN	k -nearest n eighbors
LASSO	least absolute shrinkage and selection o perator (regression)
MAPseq	m ultiplexed a nalysis of p rojections by s equencing
MERFISH	m ultiplexed e rror robust F ISH
MOp	p rimary m otor cortex
mRNA	m essenger r ibonucleic a cid
MWU	M ann- W hitney U test
PCA	p rincipal c omponent a nalysis
PT neurons	p yramidal t ract n eurons
RCA	rolling c ircle a mplification
RNA-seq	R N- s equencing
ROI	r egion(s) o f i nterest
scRNA-seq	s ingle cell R N- s equencing
sm(F)ISH	s ingle m olecule (fluorescent) <i>in situ</i> h ybridization
ST	S patial T ranscriptomics
STAR	S pliced T ranscripts A lignment to a R eference
VISp	p rimary v isual cortex

**tool names not mentioned in more than one section are not listed here*

Chapter 1

Introduction I: Biology in the era of big data

This chapter serves more as a preface to this thesis than a true introduction. For a more canonical introduction, see "Thesis Outline" in Section 1.6 and skip to Chapter 2. This chapter is adapted from an unpublished review/perspective piece, which was authored jointly by Shaina Lu, Nathan Fox, Tony Zador, and Jesse Gillis. The current draft of the full text is available in Appendix A. I wrote the original text presented here.

1.1 Neuroscience in the era of big biology

Understanding the brain is a hugely complex endeavour. This task is so grand in scope that it has inspired consortia and consortia-like efforts. In the 2000s, the Allen Institute, for example, began to release large, central databases for neuroscience. One such reference is the whole-transcriptome *in situ* hybridization atlas in the adult mouse brain and the corresponding common coordinate framework (see Section 2.4.1). The former is heavily used as a comparison for nearly all subsequent spatially-resolved transcriptomics datasets in the mouse brain and the latter provides a standard coordinate system for areas of the mouse brain (Lein et al., 2007; Ng et al., 2007). Another neuroscience consortium is the contemporary BRAIN Initiative Cell Census Network (BICCN) branch of the broader BRAIN initiative. The BICCN faction has the central goal of creating comprehensive cell type atlases across model organisms and humans. Along the way, the BICCN has

produced large transcriptional and other modality datasets, improved spatially-resolved transcriptomics tools (i.e. MERFISH, see Section 2.2.2), and more. Focusing on behavioral neuroscience, a second contemporary neuroscience consortium is the International Brain Laboratory (IBL). The IBL sought to standardize neuroscience experimental setups and behavioral tasks which are generally bespoke to individual labs in an effort to increase cross-laboratory replicability of neuroscience assays (The International Brain Laboratory et al., 2021). In the process, hardware components to assay rodent behaviour were further developed and made available.

We stand at the cusp of a potential new era for large-scale collaboration in neuroscience. While fields such as genetics and genomics now have a more than three decades rich history in consortia science, neuroscience is more newly diving deep into these big biology approaches. In 2013, Sean Eddy penned an eloquent essay pointing to successes and failures of the Encyclopedia of DNA Elements (ENCODE), largely attributing negatives to a failure to properly categorize what type of consortia ENCODE is and what its findings contribute to biology- more on this later (Eddy, 2013). Eddy ends his essay with a plea to do better for the next big consortia in neuroscience. Today, the BRAIN initiative and the International Brain Laboratory, two large consortia in neuroscience, are in full swing. With this new era of collaborative research, we must critically examine the organization, outputs, successes, and failures of past consortia to avoid past downfalls.

1.2 Biology in the era of big biology

1.2.1 Introduction to big biology research

Biology, more broadly, is in the era of big biology. Twenty years ago, at the start of the millennium, former United States President Bill Clinton and former United Kingdom Prime Minister Tony Blair jointly announced the completion of the first sequencing of the human genome (Press Secretary, 2000). It would be a few more months until the draft genome was published and a couple more years before the completion of the final version. While few scientific enterprises conclude with the fanfare of the Human Genome

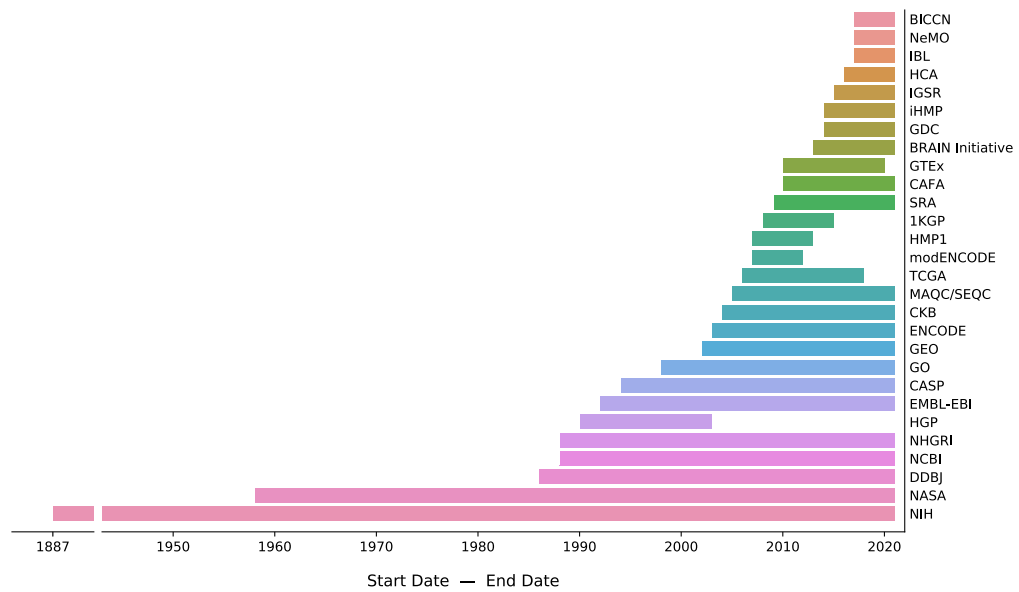


FIGURE 1.1: **Lifespan of a selected group of consortia and consortia-like research initiatives.** Plot showing the start and end date (x-axis) of a selected group of consortia and consortia-like research initiatives (y-axis).

This figure was created by Nathan Fox.

Project (HGP), this moment marked a milestone for all biological and biomedical research. Not only would the research findings of this consortium become foundational for modern research in these fields, the technology developed, standards- formal and informal- set, and style of research would transform modern biology. Notably, prior to the conclusion of the HGP, there were a handful of sometimes overlapping consortia organized around sequencing the comparatively smaller genomes of model organisms (The C. Elegans Sequencing Consortium, 1998; Bowman et al., 1997; Adams et al., 2000). Together, these genome sequencing consortia represented the start of consortia-based big science in biology.

Today, in the nearly two decades since the HGP announcement, consortia of all types permeate all corners of biology (Figure 1.1). These big science approaches have generated previously unthinkable datasets: the Allen Institute’s brain-wide spatial expression atlases, the extensive characterization of DNA by ENCODE, and the organized functional gene annotations of GO- to name a few. They have set forth laudable data sharing and ethics principles: the public data sharing guidelines of the HGP’s Bermuda Principles and the establishment of the Sequence Read Archive (SRA) in support of the 1000 Genomes

Project and Human Microbiome Project. They have democratized tools for research: sequencing technologies from the HGP and STAR, the sequence alignment algorithm from ENCODE. Beyond these tangible contributions to science, consortia also provide inspirational value to humanity by increasing public trust in science and promoting cross-cultural, international collaborations.

Despite all of these contributions to biology, the consortia-based approach is not the be-all and end-all of research. As with any large organizations, consortia can promote group mindsets and hamper creativity. Smaller, dynamic groups may find it hard to compete with resource-rich consortia further propagating the Matthew's effect present in science (Merton, 1968; Bol, Vaan, and Rijt, 2018). Matthew's effect is often defined via the saying: "the rich get richer and the poor get poorer." Further, consortia often focus on better-established, mainstream research topics, leaving newer, unique ideas to smaller groups (Bhattacharya and Packalen, 2020). Research on these unique ideas has proven pivotal in the past. (Famously, the discovery of Taq polymerase came from research on the Yellowstone hot springs.) Additionally, trainees involved in large-scale collaborations may find it difficult to get the recognition necessary for scientific career advancement. These are just a few of the possible concerns with big biology.

What constitutes a consortium or not in biological research is not black and white and can vary between different researchers. For the purposes of this chapter, we define the label consortium very loosely according to Merriam-Webster as "an agreement, combination, or group (as of companies) formed to undertake an enterprise beyond the resources of any one member." Under this umbrella we include discussion here of many multi-lab research efforts that may not be considered a consortium by most scientists since they are competition-based. In general, we limit our discussion of consortia to North American groups in the fields of genomics and neuroscience given the bias in the authors' location of work and experience.

1.2.2 Diversity of consortia

Even before modern-day consortia, humans have long been organizing themselves behind common scientific goals. During the age of exploration and beyond, determining a ship's location on long ocean voyages was key. While latitude was easily found by tracking the sun, longitude was a notoriously hard problem. To incentivize this, prizes were offered by European rulers as early as the mid-16th century and as recently as the longitude rewards of the British government established in the early 18th century. Also in the 18th century, when Italy was still a bunch of fragmented states, scientists across what would later become a unified Italy, decided they wanted to find eel gonads and solve the mystery of their reproduction (*Slippery Mystery* | Radiolab 2020). The question of eel reproduction was so perplexing and long-standing, that they believed this discovery would be a part of an unifying Italian national brand. A more recent example of big-science is the space race of the US and Russia in the cold war era. In the US, the multi-state and multi-billion National Aeronautics and Space Administration (NASA) was created. Incentivization, collaboration, and competition through large-scale scientific efforts has been a part of human civilization for centuries.

In biology today, consortia come in all shapes and sizes. These consortia vary not only in their research focus, but also in how they are created, organized, and funded—to name a few axes of variation. Take for example, the Critical Assessment of Structure Prediction (CASP). CASP is a long-running (biannually since 1994), grassroots collection of scientific groups working toward the common goal of predicting unknown protein structures with computational modeling. At its core, it is a contest where the individual groups work in competition with each other to build the best structure predictors. (This competition has been in the news lately because of the complete domination of its central task by DeepMind's AlphaFold2 (AlQuraishi, 2020).) Notably, there are also a number of other competitions under the umbrella of DREAM Challenges organized to address computational problems across biology (Ellrott et al., 2019). It should be explicitly stated, that these competitions are not consortia in a traditional sense, but can be viewed under the

larger umbrella of multi-lab, big biology research with a unified goal; though the mechanism is direct (friendly) competition rather than direct cooperation. (In the field of structure prediction/structural genomics, there were a variety of more traditionally cooperative consortia (Todd et al., 2005).) In contrast to these competition-based initiatives, are highly-centralized consortia, such as the previously mentioned BICCN arm of the broader BRAIN initiative. The BRAIN initiative came about as a project launched and funded through a US Presidential initiative of the Obama administration. The BICCN faction has the central goal of creating comprehensive cell type atlases across model organisms and humans. This consortium is a goal- and funding- driven initiative very dissimilar to CASP.

A third class, tangential to both CASP- and BICCN-like consortia, are consortia organized around longitudinal data collection. A couple of famous examples of this are the UK Biobank and Framingham Heart Study. These types of consortia run over long periods of time to collect longitudinal data that would likely not be possible without the respective consortia's existence. The Framingham Heart Study has been running since 1948 in its namesake town of Massachusetts (Andersson et al., 2019). This study now includes over 14,000 participants spanning three generations and is responsible for much of our modern-day understanding of the risks and prevention of cardiovascular disease (Mahmood et al., 2014). Not unlike BICCN, the Framingham Heart Study was started through the National Heart Act signed by former U.S. President Harry Truman in 1948 (Mahmood et al., 2014). (Truman was vice president to Franklin D. Roosevelt who suffered from largely undiagnosed cardiovascular disease.) While younger in age, the UK Biobank is also a longitudinal study. Started in 2006, the roughly 500,000 participants agreed to be followed for at least 30 years (Bycroft et al., 2018). The scope of this data is extensive ranging from simple survey demographics, MRIs of the brain and heart, and genotyping of blood samples. While some of this data collection is still on-going, this dataset has already proved invaluable. Presently, the UK Biobank resource has also enabled researchers to name putative risk factors of COVID-19 (Armstrong et al., 2020; Yates et al., 2020) and to identify race and socioeconomic demographics of COVID-19 infections (Niedzwiedz et al., 2020). These large-scale data-based consortia enable well-supported, population-level

studies both in the world that conceived the studies and well beyond. They enable rapid research and understanding in real-time crises.

1.3 Consortia have enabled massive data collection and tool development

1.3.1 Data

In modern biology, with the competition/benchmarking based groups as an exception, one of the more ubiquitous outputs of research consortia is data. The large production scale of consortia has enabled massive data collection and tool development. The flagship output of the HGP is the reference human genome which expanded into cataloguing all functional elements of the genome in ENCODE and model organisms in modENCODE. As sequencing prices dropped, many consortia sprung up around cataloguing the genomic diversity of, often previously underrepresented, human populations: the 1000 Genomes Project (now, The International Genome Sample Resource) sequencing individuals across the world in an effort to identify rare variants, UK10K in sequencing 10,000 UK individuals to identify rare variants, GenomeAsia 100K Project in wanting to add diversity to genome datasets by sequencing across Asia (GenomeAsia100K Consortium, 2019), and perhaps most recently, the All of US precision medicine group from the National Institutes of Health (NIH) seeking to gather biological and health data from over 1 million U.S. participants. Similarly, data-focused initiatives also grew around specific interest areas of biology: The Cancer Genome Atlas Project (TCGA) sought multi-omic profiling of 20,000 cancer samples with matched healthy samples, the Allen Institute (itself more of a research institution than consortia) created a comprehensive in situ hybridization based spatially-resolved transcriptomic atlas of the developing and adult mouse brains (among many other atlas style resources), and the Genotype-Tissue Expression (GTEx) consortia matching genotype with tissue specific transcriptomics. As technologies were developed and improved, additional data-generating groups utilized these new tools: single cell mouse brain atlases of the BICCN and Allen Institute and the tissue specific single cell atlases of

the larger Human Cell Atlas (HCA) and Chan Zuckerberg Initiative's (CZI) Tabula Muris. Finally, it is worth re-highlighting as previously discussed that longitudinal datasets requiring multi-generational organization are mostly impossible outside of consortia. (Richard Lenski's directed evolution experiment in bacteria is a major exception.) These examples are by no means all encompassing, but do illustrate the broad focus on data for many consortia.

Concerning human data, there is a notable lack of diversity in these datasets. For example, not only are the subjects of the Framingham Heart Study predominantly white and of European origin, but this demographic make-up was further claimed to be representative of the 1940s U.S. when the study started (Mahmood et al., 2014), which is simply not true. GTEx, as a recent example, is 84.6% white and 67.1% male (GTEx Portal) which is not representative of the racial and gender make-up of its host country, the US. Lack of diversity in datasets is not only harmful to the communities that these consortia fail to serve, but also harmful to the research itself. Concretely, diverse human datasets would for example allow researchers to identify more polymorphisms and more generally allow for more robust and generalizable research findings. Relatedly, there has been a recent resurgence in improving the human reference genome to be more representative of the human population and not just consisting of mostly one single individual as the dominant reference is today. One solution is a consensus genome that would be able to harness the diversity in sequenced genomes to build a better reference (Ballouz, Dobin, and Gillis, 2019). Recent consortia promise to increase the diversity of human datasets such as the GenomeAsia 100K project, All of Us, and the HCA. As the vanguard for large-scale data generation, consortia are continually responsible for ensuring the diversity of their human datasets.

1.3.2 Data sharing

In many cases, some of these large-scale and/or long-scale datasets likely only exist because of consortia-like effort. However, relative to other highly collaborative fields,

high-energy physics for instance, biological data is extremely fragmented. With the diversity of data types and constantly evolving technologies, data and associated meta-data in biology is extremely messy, often with inconsistent formatting. In efforts to combat this, there are laudable data-sharing policies and infrastructures that grow out of consortia. Leading the way in data sharing, prior to consortia even, was the Protein Data Bank (PDB) which was the first database of its kind in biology (“Crystallography” 1971). The PDB was an open access repository to freely share protein structures with the community. Following these traditions, the HGP consortium also led in open science by requiring sequence data to be rapidly released prior to publication as laid out in the Bermuda Principles. This set the standard for genomics research, which continues to be one of the most open sub-fields of biology; the genomics field was one of the earliest adopters of bioRxiv pre-printing, having some of the highest numbers of pre-prints (Sever et al., 2019). In the early 2000s, out of community demand, the National Center for Biotechnology Information (NCBI), a branch of the NIH, established the Gene Expression Omnibus (GEO) for the main purpose of sharing gene expression data in the form of microarrays (Edgar, Domrachev, and Lash, 2002). Impressively, GEO continues to be a major resource for sharing multi-omic data today. As previously mentioned, later in the 2000s, the NCBI, in collaboration with the European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ), established the Sequence Read Archive (SRA) in support of international consortia: the Human Microbiome Project and 1000 Genomes project (Kodama, Shumway, and Leinonen, 2012). The SRA is still used today as one of the main repositories for sequencing data. Recently, as a part of the BRAIN Initiative, the Neuroscience Multi-omic Data Archive (NeMO) was introduced as a multi-modal data repository for modern neuroscience datasets spanning physiology, genomics, and beyond.

These days, however, as datasets continue to proliferate in both quantity and individual size, data availability is often not enough to render them useful to the research community. One barrier to accessibility is poor metadata, a problem recently compounded by single cell sequencing platforms which GEO was not designed to support. Recognizing

this, the BICCN, for example, is currently reckoning with how to make their data continually accessible and easy to use. A part of the BICCN collaboration, the Karchenko Lab has proposed a Cell Type Annotation Platform (CAP) to standardize the sharing of single cell data and associated metadata (*cap-example* 2020). Also going further than simply serving data, there have been recent efforts to pre-process large bodies of data using the same bioinformatic pipelines such as the Genome Data Commons (GDC) of the National Cancer Institute (NCI), which has consistently processed data across many large cancer genome datasets including the aforementioned TCGA (Heath et al., 2021).

Beyond the influence of consortia efforts, there are additional concerns and barriers to data sharing. With human data, ensuring the safety and privacy of donors is paramount. Sometimes, however, data protection regulations can inadvertently inhibit scientific data sharing. One example is the European Union's General Data Protection Regulation (GDPR) which was passed in 2016 and went into effect in 2018. While the target of the regulation was to protect personal data, genomics data on human subjects can sometimes fall under this regulation making it challenging for genomic and health data sharing both within and beyond the EU (Eiss, 2020; Molnár-Gábor and Korbel, 2020). In the two years since going into effect, frustration in the GDPR's lack of clarity in interpretation abounds and international collaborations, including consortia, have been stalled (Eiss, 2020; Molnár-Gábor and Korbel, 2020). Beyond the EU, researchers who want to work collaboratively on human data across international borders may need official appointments or contracts drawn to even access the data. While these protections on human data are ultimately good, setting up new collaborations can become costly much like the cost of setting up new partnerships on the open market. Having established consortia, like firms, can help alleviate these barriers to working collaboratively on human data.

1.3.3 Democratizing new technologies and computational software

Being at the forefront of data generation means that consortia are often also leaders in the creation and democratization of technological development. In order to achieve the goals of a consortium, new technology needs to be pushed to production scale in contrast

to the proof-of-principle style of individual labs. Again, the HGP provides a touch stone example, in that sequencing of the human genome started out using the laborious bacterial artificial chromosome sequencing (BAC-sequencing) technique where fragmented human DNA was cloned into bacteria for replication then sequencing and re-assembly. By the end, through Craig Venter's group that split off, the human genome was competitively sequenced using shotgun sequencing, a more efficient sequencing approach that eliminated the bacterial cloning step and laid the foundation for future sequencing. The development and commercialization of production-scale sequencers can be viewed as a direct outgrowth of the HGP. A more recent example is the previously mentioned IBL consortium. In behavioral neuroscience, research methods including protocols, hardware setup, and behavioral tasks themselves often vary widely between different research groups. IBL sought to use one behavioral task with standardized equipment and protocols across various labs to determine the replicability of their assayed results (The International Brain Laboratory et al., 2021).

These technological advancements are not limited to hardware; many popular scientific software have grown out of consortia as well. For example, the popular RNA sequencing read alignment algorithm Spliced Transcripts Alignment to a Reference (STAR) was developed to align reads generated from ENCODE (Dobin et al., 2013). STAR is now widely taught and used in bioinformatics courses and research, respectively. In keeping with the advancement of sequencing platforms, subsequent versions of STAR for single-cell RNA sequencing have been developed (Blibaum, Werner, and Dobin, 2019). A second example is the SpaceTx pilot project of the HCA which sought to benchmark and streamline the analysis of spatially-resolved transcriptomics. This resulted in the Starfish suite of tools to analyze spatial expression datasets. The SpaceTx effort even included a hackathon to bring together the community in working on this tool (Perkel, 2019).

In many ways, outside of commercialization, the maturation of technologies in the academic realm is only possible through consortia efforts. Outside of consortia, funding to develop technology beyond a prototype is scarce in current science funding models. Consortia can provide the research dollars needed to mature a technology. With so many

bespoke technologies in individual labs, one way to think of the relationship between consortia and technology development is in reference to economic theories on international trade and economies of scale (*The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2008*). Grossly simplifying, economies of scale states that it is cheaper to mass produce a product. Generally, mass production would reduce individual prices, but global trade mediates this. We can abstract this further as a hub and spoke model, where in consortia we have a hub, either a single lab or whole consortia, that has the expertise to run a complex technique and provide it as a service to other research groups, the spokes. Another phenomenon related to consortia and technology worth noting, is that sometimes the adoption of a technology by a consortium can cause a competing approach to be sidelined for little apparent reason.

1.4 Robustness of new technologies is critical

With the inundation of big data and new technologies through consortia, we now have the opportunity to assess the robustness and replicability of these datasets and technologies. One contribution of consortia in ensuring the robustness of research is through setting research standards. Having power in numbers, and often funding, consortia have the leverage to do so. An obvious example of a research standard is the human reference genome first published by the HGP. The importance of having an accepted standard reference in genomics research can be stated by analogy to the need to have a reference for a standard measurement, much like the recently redefined official kilogram reference (Stock et al., 2017). Beyond providing a reference, the entire purpose of some consortia can be to define research standards. For example the MicroArray or subsequent Sequencing Quality Control (MAQC/SEQC) effort from the U.S. Food and Drug Administration (FDA) is entirely organized around benchmarking transcriptomics technologies so that they could be reliably used in diagnostic and regulatory applications (Canales et al., 2006; Shi et al., 2006). A second, slightly different, example of a research standard is the controlled vocabulary to define gene function created by the Gene Ontology (GO) (Ashburner et al., 2000;

The Gene Ontology Consortium et al., 2021). With the tagline goal of “Unifying Biology,” GO has enabled standardized annotations of gene function and the easy assignment of functional enrichment for any given gene. Yet another example of ensuring robustness in biology research is the PDB, which by providing a database for depositing protein structures has allowed researchers to iterate on them and further increase the resolution of the structures. Whether an offshoot of data generation or the central goal of a consortium, consortia often play a large role in defining and setting research standards.

Outside of consortia themselves, since there is so much data available and multiple techniques to assay the same thing, it is the collective responsibility of the research communities to independently benchmark these datasets and technologies (Boulesteix, Lauer, and Eugster, 2013) (see Section 2.5). In an ideal world, we should be able to compare data from different technologies that are assaying the same biological phenomenon (i.e. different microarray platforms). The availability of these independent benchmarks would allow us to understand the pros and cons of various methods and decide which technologies to further develop in a fair way.

1.5 Conclusions

Writing in *Science Magazine* in 1961, before the era of biological consortia, Alvin Weinberg, then director of the Oak Ridge National Laboratory, drew an analogy to the big science projects of the day to historical monuments like the Egyptian pyramids, the cathedrals of the middle ages, and the Palace of Versailles even warning about the link between these projects and the demise of the economies that conceived them. Alvin argued that big science could seriously harm scientific research as a whole, leading to the triple diseases of “journalitis, moneyitis, administratitis” (Weinberg, 1961). In closing, he argued that we should focus our efforts on improving human well-being. Today’s biological consortia arguably do just that. Half a century later, in 2012, the authors and participants of a review on consortia efforts in immunology, refuted Weinberg, writing: “But, surely, finding a fundamental particle of the Universe or deciphering the human genome has inspirational value at the individual and societal level that transcends any usual science project”

(Benoist et al., 2012). Consortia have the potential to transcend economic analyses and provide inspirational value.

As discussed above, consortia science can democratize tools for science by making them more widely available and affordable, promote open science through collaboration, and lead to the growth of a nation. Even beyond these benefits, as a large-scale effort, consortia have the potential to capture public attention. Some consortia (i.e HGP) have the star power to help shape the public narrative and responsibly foster public trust in science. Trust in science is ultimately a good thing (perhaps we may have better bore the early brunt of COVID-19, for example). At times also inherent to their large-scale nature, consortia often represent collaborations across countries and cultures. Not only can this accelerate data access as previously mentioned, these collaborations can bridge across geopolitical boundaries. While not a consortium, during the cold war, for example, the U.S. and the Soviet Union collaborated on bringing a second Polio vaccine to market. Albert Sabin developed a polio vaccine using attenuated polio virus, but it could not be tested in the U.S. since an earlier vaccine developed by Jonas Salk was in use. In collaboration with Soviet scientists Mikhail Chumakov, Maria Voroshilova, and Anatoli Smorodentsev, Sabin's vaccine was able to be tested in the USSR which had active polio outbreaks, proved efficient, and ultimately used for vaccination of children in both countries (Horstmann, 1991). Today, many consortia bridge international boundaries, providing opportunities to scientists from various backgrounds, increasing cultural competency, and ultimately strengthening science itself through the diversity of its participants.

1.6 Thesis Outline

One rapidly developing technology is spatially-resolved transcriptomics. Some of these tools are being further developed as part of a consortia (e.g. MERFISH, though pre-dating the consortium, was further developed within BICCN (see Section 2.2.2)), while other new approaches are coming out of individual labs. Additionally many companies have jumped into the game too (e.g. Spatial Transcriptomics developed as Visium through

10x Genomics (see Section 2.2.1)). Regardless of the entity developing these tools, spatially-resolved transcriptomics is a field packed with emerging experimental technologies ripe for replicability analysis. There are many opportunities for data analysis, building tools, and integration of data from different technologies with the increasingly wide adoption of spatially-resolved transcriptomics in biology today.

In Chapter 2, I introduce the field of spatially-resolved transcriptomics more generally, outlining experimental assays, computational analysis, and applications to multi-modal neuroscience. I conclude Chapter 2 with a discussion on the need for replicability assays within this emerging field. In Chapter 3, I perform one such replicability study between Spatial Transcriptomics (Ståhl et al., 2016) and *in situ* hybridization (Ortiz et al., 2020; Lein et al., 2007), benchmarking two whole-brain, whole-transcriptome datasets collected in the adult mouse brain using each of these technologies. While replicability is not always one-to-one, biological conclusions from these two datasets generally replicate. Following this, in Chapter 4, I share three vignettes on more specific spatial applications in neuroscience. First, I assess the replicability of the expression of the cadherin gene family in the primary auditory cortex between *in situ* sequencing (BARseq) and *in situ* hybridization (ABA). Next, I ask if patterns of spatial expression are driven by cell type composition by linking spatially-resolved expression with single-cell data. Lastly, in Chapter 4, I create a pipeline for base-calling *in situ* sequencing (BaristaSeq) reads of random barcodes for projection mapping. Finally, Chapter 5 is the conclusion, summarizing and providing some perspectives on our results more generally.

Chapter 2

Introduction II: Spatially resolved transcriptomics

2.1 Spatially-resolved transcriptomics is poised to be transformative across biology

The recent explosion and increased accessibility of single-cell RNA-sequencing techniques has been transformative to studying gene expression across biological questions. As single-cell techniques mature, one central limitation that has become obvious is the inability to assay the spatial origin of the cell within the tissue in conjunction with gene expression. Within the last 5 years, new spatial transcriptomics techniques have made it possible to link expression with the spatial origin of transcripts in a high-throughput manner. These new spatial technologies are poised to be transformative, earning the distinction of Nature Methods 2020 Method of the Year (Marx, 2021).

Until recently, spatially resolved expression data was laborious to obtain and generally low-throughput. Recent innovations have introduced improvements on older methods, such as increased multiplexing of *in situ* hybridization (ISH), or represent entirely new technologies (see Section 2.2.1). These new methods are highly accessible and readily adopted across biological fields, including non-model organisms and disease applications (Giacomello et al., 2017; Lundmark et al., 2018). Like most widely adopted sequencing methods, these methods are tissue agnostic, allowing wide application with minimal

species-specific optimization. Recent and continuing commercialization of new spatial methods has further democratized access to spatially resolved transcriptomics.

Finally, the availability of spatial information in assaying gene expression allows researchers to answer a whole new class of biological questions. The most obvious is the ability to characterize spatial patterning of expression in tissues (e.g. Halpern et al., 2018; Moor et al., 2018; Codeluppi et al., 2018), organisms (Chen et al., 2021a), and even colonies of single-cell organisms (Vliet et al., 2018). Further, application to developing specimens can identify spatial gradients of expression, key to development. Spatial expression is poised to link molecular properties, such as expression, with macro-scale features such as cytoarchitecture and tissue structure (Lein, Borm, and Linnarsson, 2017). The interactions between cell-types identified from single-cell biology can be observed for the first time in their native organization. Single-cell biology was the revolution of the last decade; spatial has the potential to be the revolution of this decade.

In this chapter, I first introduce various types of experimental spatially resolved assays (Section 2.2), breaking them down into capture-based, *in situ* hybridization, *in situ* sequencing, microdissection, and others (Table 2.1). Next, I discuss the introduction of computational methods to analyze spatial data (Section 2.3). I attempt to organize the computational tools according to their primary tasks, but it is clear that they often overlap (Table 2.4). I then focus on the role of spatial expression techniques in neuroscience as a key part of multi-modal neuroscience studies (Section 2.4). This section is split into details about early spatially-resolved transcriptomics in neuroscience (namely, the Allen Brain Atlas), current examples of multi-modal neuroscience studies, and the development of experimental tools of multi-modal neuroscience. Finally, I conclude with a discussion on the necessity of benchmarking of spatially-resolved techniques, especially necessary with all these newly developed methods (Section 2.5).

2.2 Spatial techniques, experimental

There are a variety of experimental approaches to assay spatially resolved expression (2.1) (Regev et al., 2018; Asp, Bergenstråhle, and Lundeberg, 2020). The oldest of

TABLE 2.1: Comparison of types of spatially resolved transcriptomics approaches.

Technique Class	First Demonstrated	Whole Transcriptome or Targeted	Readout
Capture-based	2016	Whole-Transcriptome	Sequencing
<i>in situ</i> hybridization	1982	Targeted	Imaging
<i>in situ</i> sequencing	2013	Targeted (except FISSEQ)	Imaging
microdissection	1996*/2013	Whole-transcriptome	Sequencing

*LCM published (Emmert-Buck et al., 1996)

these methods are ISH-based methods, with single molecule *in situ* hybridization dating back to the early 1980s (Asp, Bergenstr hle, and Lundeberg, 2020). In recent years ISH methods have been further developed with multiple technical improvements along side the introduction of *in situ* sequencing methods (ISS). Most recently a whole new class of spatially resolved assays have emerged, loosely referred to as capture-based approaches. This class of methods is so named because of the use of spatially-barcoded probes to capture (generally) transcriptome-wide messenger RNA (mRNA) transcripts (see Section 2.2.1). Outside of these three classes, there are additional spatial expression methods such as microdissection followed by traditional RNA-sequencing (RNA-seq). In this section, I will focus on capture based approaches, as one of the most innovative sub-groups of spatially resolved transcriptomics, while more briefly introducing some of the others.

2.2.1 Capture-based methods

Capture-based methods are a class of spatial transcriptomics methods relying on mRNA probes. Generally these probes contain a poly-T region to hybridize against the poly-A tail of mRNA transcripts along with a unique spatial barcode (Figure 2.1). The patterning of these spatially barcoded poly-T probes is known either by design or are read out

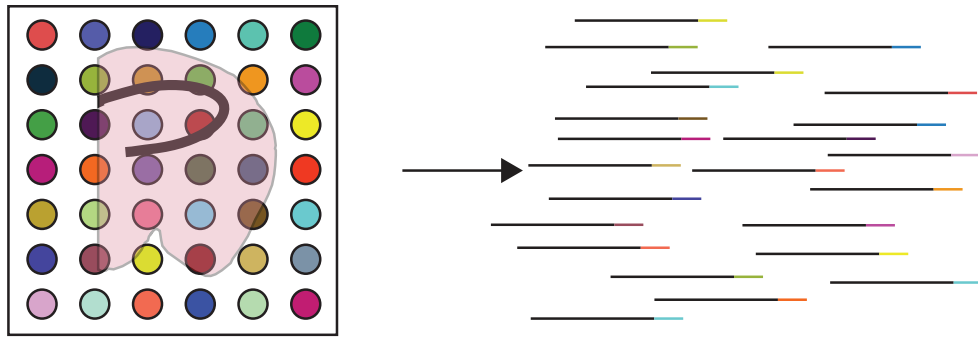


FIGURE 2.1: **Simplified schematic of most capture-based spatially-resolved transcriptomics methods.** Left: A schematic representing a slide with spots containing unique barcodes (represented by colored spots here) that hybridize to mRNA from the example tissue slice placed on it. The space between each unique spot varies by method, with some having essentially negligible space unlike this depiction. Right: After pre-processing and sequencing, mRNA reads include an associated spatial barcode representing the spatial origin of the transcript.

using an initial hybridization or sequencing step prior to mRNA hybridization. By targeting the poly-A tail, these capture-based methods are able to target the whole transcriptome without the need to design libraries of probes specific to each gene in the genome (as most ISH and ISS require). After mRNAs are captured and spatially barcoded, read out of spatial gene expression is done using sequencing. Capture-based methods are high-throughput, easily spanning the whole transcriptome. With each new technology, the spatial resolution of the capture-based approach continues to drop (Table 2.2). However, the increase in spatial resolution does not always track with an increase in transcript capture (Table 2.3). In fact, in some cases transcript capture even decreases with higher spatial resolution requiring these higher resolution spots to be binned together into lower resolution spots for analysis (Vickovic et al., 2019; Rodrigues et al., 2019). This class of spatial transcriptomics methods have proven to be particularly advantageous in modern research.

ST and 10x Visium. The first capture-based method developed is Spatial Transcriptomics (ST) published in 2016 (Ståhl et al., 2016). It was a ground-breaking proof-of-principle, opening the doors to the spatial revolution. ST works by tiling spatially barcoded poly-T probes on a glass slide. Conceptually, this works much like a microarray, only the probes in each spot are not targeting specific genes, but all mRNA and each probe spot has

TABLE 2.2: Comparison of spatial resolution given by spot density in capture-based spatially resolved transcriptomics techniques.

Technique	Reported Spatial Resolution (Spot Density)
ST (Ståhl et al., 2016)	100 μm spots, 200 μm center-to-center
Visium (10x Genomics)	55 μm spots, smaller distance than ST
HDST (Vickovic et al., 2019)	2 μm binned to 13 μm for analysis
Slide-seq (V1/V2) (Rodriques et al., 2019; Stickels et al., 2020)	10 μm beads
DBiT-seq (Liu et al., 2020)	10, 25, or 50 μm width channels
stereo-seq (Chen et al., 2021a)	220 nm (0.22 μm) diameter, 500-715 nm (0.5-0.715 μm) center-to-center
Seq-scope (Cho et al., 2021)	0.5-0.8 μm apart spots
PIXEL-seq (Fu et al., 2021)	$1.17 \pm 0.1 \mu m^2$ spots
sci-Space (Srivatsan et al., 2021)	single-cell barcoded by $\sim 73 \mu m$ radius spots, $\sim 220 \mu m$ center-to-center

TABLE 2.3: Comparison of capture sensitivity in capture-based spatially resolved transcriptomics techniques.

Technique	Detection Efficiency
ST (Ståhl et al., 2016)	6.9%*
Visium (10x Genomics)	higher than ST*
HDST (Vickovic et al., 2019)	1.3%*
Slide-seq (V1/V2) (Rodriques et al., 2019; Stickels et al., 2020)	V1: 0.3%* or ~ 300 -1000 total transcripts/total cells; V2: 10x better
DBiT-seq (Liu et al., 2020)	~ 4000 unique molecules per 10 μm spot
stereo-seq (Chen et al., 2021a)	~ 7 transcripts per 4 μm^2
Seq-scope (Cho et al., 2021)	~ 6 (liver) - 23 (colon) unique molecules per $< 1 \mu m^2$
PIXEL-seq (Fu et al., 2021)	> 100 unique molecules per 10x10 μm^2
sci-Space (Srivatsan et al., 2021)	2514 unique molecules per cell

*Asp, Bergenstråhle, and Lundeberg, 2020

a unique spatial barcode. Fresh frozen tissue (and now also fixed tissue for 10x Visium) is placed on the barcoded slide where mRNAs can hybridize to these probes. Captured probes on the glass slide then undergo reverse transcription to form complementary DNA (cDNA) which can then be removed from the slide and sequenced using normal next-generation RNA-seq. These sequenced transcripts contain a spatial barcode from the probes that allow detected transcripts to be mapped back to their spatial location of origin. The resulting dataset is a count of 3' cDNA of each transcript per spot on the slide.

In the last couple of years (roughly 2019), 10x Genomics launched a commercialized version of this method called Visium. The Visium approach uses the same conceptual approach, but increases the spatial resolution (Table 2.2) and transcript capture rate through optimization of ST. Capture rate is reported as the number of unique molecules (transcripts) or genes detected per spot or cell. Visium is now readily available with downstream bioinformatic analysis supported through freely available company software, Space Ranger and Loupe. It has already been implemented in many labs (e.g. Maynard et al., 2021).

HDST. High-Density spatial transcriptomics (HDST) was published in 2019 (Vickovic et al., 2019) by many of the same authors as the ST method. HDST greatly improves the spatial resolution of ST (Table 2.2) by using a similar barcoding strategy on beads packed into a 2-dimensional plane. Since the distribution of barcoded beads on the array is unknown, spatial barcodes must first be read-out through, here, hybridization prior to assaying the sample. While the resolution is greatly improved, capture rate is low requiring neighboring beads to be binned together into lower-resolution spots for downstream analysis.

Slide-seq V1/V2. Slide-seq is very similar to HDST. V1 was shared on bioRxiv just before HDST (Rodrigues et al., 2019). Slide-seq uses $10 \mu m^2$ barcoded beads similar to those used in single-cell RNA-sequencing (i.e. Drop-seq) and packs them into what the authors call a 'puck' on a glass slide (Rodrigues et al., 2019). Each of these beads contains a unique barcode that must be read out by sequencing to give the spatial positioning of the bead on the slide. Subsequent steps are very similar to ST. Though the spatial resolution

of slide-seq V1 is high, gene detection is low. Slide-seq data required the input of single-cell RNA-sequencing (scRNA-seq) data for downstream analysis. Slide-seq V2 reports the same spatial resolution, but improved chemistry for 10-fold better transcript capture (Stickels et al., 2020).

DBiT-seq. Published in late 2020, deterministic barcoding in tissue for spatial omics sequencing (DBiT-seq) diverges from the above methods by using a microfluidic array to directly tile tissue with spatial barcodes (Liu et al., 2020). The method is based around a microfluidic chip of 10, 25, or 50 μm width channels which applies known spatial barcodes to the tissue. Following reverse transcription, the chip is then rotated by 90° and spatial barcodes are applied in this orientation as a second step. Barcodes from the second step are ligated to the first-step barcodes. The result is a known spatially barcoded grid pattern on the cDNA in the tissue which can then be extracted for sequencing.

Notably, DBiT-seq can apply spatial barcodes ligated to oligos for hybridization with mRNA or barcodes attached to antibodies to study spatially-resolved proteomics. At publishing, this technique was one of the more multi-omic friendly approaches. This was illustrated with two follow-up papers from the same research group demonstrating spatially-resolved chromatin profiling combined with DBiT-seq including Cut&Tag for assaying histone modifications and ATAC-seq for chromatin accessibility (Deng et al., 2021b; Deng et al., 2021a).

Stereo-seq. Spatio-Temporal Enhanced REsolution Omics-sequencing, or Stereo-seq, was shared as a pre-print in early 2021 (Chen et al., 2021a). Stereo-seq is similar to the above capture-based methods, using probes to capture transcriptome-wide mRNAs. Stereo-seq improves on the spatial resolution and capture, by using DNA nanoballs with random barcode sequences that can 'dock' on an etched silicone chip. The barcodes on this chip are read-out using sequencing. Then poly-T probes are ligated to the nanoballs creating mRNA probes. Tissue is placed on the chip, followed by hybridization, reverse transcription, and sequencing. The advantage of the DNA nanoballs is the tight-packing for high spatial resolution combined with longer barcodes allowing for more unique probes. Stereo-seq has the highest reported spatial resolution (Table 2.2).

Seq-scope. Seq-scope, shared as a pre-print in early 2021 and published in mid-2021, takes advantage of Illumina sequencing to perform spatially-resolved transcriptomics assays directly on an Illumina flow cell (Cho et al., 2021). Reporting a spatial resolution of 0.5-0.8 μm center-to-center, seq-scope has the highest resolution of published methods (not including some comparable or better pre-prints) (Table 2.2). Seq-scope works through two sequencing steps; the first step generates the spatial barcodes on the flowcell, and the second step reads out the hybridized transcripts. The spatial barcodes contain PCR adapters that bind to the flowcell and are amplified to create spots containing the same barcode. Taking advantage of Illumina sequencing-by-synthesis the location of the spatial barcodes are read out without requiring any bespoke analysis. Cutting the barcodes in a region of the oligo designed to be recognized by an enzyme, the hybridizing region is then exposed. Now, the second sequencing step commences in a manner similar to ST, HDST, and slide-seq. Fresh frozen tissue can be placed on the flowcell allowing mRNA to hybridize to the probes created in the first sequencing. A hybrid probe-cDNA oligo is synthesized and sequenced.

Seq-scope also reports a high capture rate, having a higher capture rate per pixel than ST, HDST, slide-seq, and DBiT-seq (Cho et al., 2021). Processing and downstream analysis of seq-scope data notably uses standard bioinformatics tools (Illumina Real-Time Analysis, STARsolo, and Seurat). The high spatial resolution, high capture, and standard analysis of seq-scope combine with relative cost and time efficiency (two rounds of Illumina sequencing), to make seq-scope an extremely appealing tool for wide adaptation. Efforts are currently underway for commercialization.

PIXEL-seq. Polony-indexed library-sequencing (PIXEL-seq), similar to seq-scope, is iterating on Illumina sequencing technology for spatially-resolved transcriptomics. PIXEL-seq was shared on bioRxiv in 2021 a few months after seq-scope and reports a similar, though very slightly lower, spatial resolution to seq-scope (Table 2.2) (Fu et al., 2021). Traditional sequencing surfaces tend to have large physical peaks on the surface, so PIXEL-seq modifies this using a polyacrylamide gel optimized to have a continuous surface for

best mRNA capture. Spatial barcodes are amplified on these custom gels cast into an Illumina-compatible flowcell and then sequenced using a custom sequencing-by-synthesis set-up. Fresh frozen tissue is then applied to this gel allowing hybridization of probes to mRNA, reverse transcription, and sequencing. Details aside, this approach is synonymous to seq-scope, but requires more bespoke chemistry and tools.

Sci-Space. The newest of the capture-based approaches, Sci-space was published in mid-2021 (Srivatsan et al., 2021). Though no pre-print had been available, it was discussed at various scientific conferences in the past couple of years. Much like the early capture methods, sci-space uses a barcoded array, here with 2 spatial barcodes, read out through hybridization, to which a fresh frozen tissue slice can be applied. In contrast, in a sense, sci-Space flips the conventional capture-based workflow on its head by having barcode oligos be taken up by the cell instead of lysing cells for mRNA hybridization. The idea of using barcode oligos in this way grows from some of the same authors' previously published scRNA-seq method (Cao et al., 2017). Since the cells are intact, the cells can then be dissociated from the slice and tissue and mRNA in the cells can then be labeled with an additional slice barcode oligo and nuclei barcode, following the scRNA-seq protocol (Cao et al., 2017). Briefly, this scRNA-seq protocol labels fixed single cells by barcoding each well of cells distributed in a 96-well plate. The same cells are then pooled and re-distributed by fluorescent sorting into a second 96-well plate where they are barcoded a second time. After sequencing, the combination of these added barcodes are then used to assign spatial and cell identity to the reads. In this manner, sci-space is the first spatially-resolved method with single-cell resolution similar in theory to that of scRNA-seq (barring doublets, etc.). Though it should be noted that the experimental combination of scRNA-seq on cryosections is hard and can often affect sequencing quality.

Other capture methods. There are a handful of other spatially resolved transcriptomics methods that are sometimes classified as capture-based methods (Asp, Bergenstråhle, and Lundeberg, 2020). For example, the GeoMx Digital Spatial Profiling (DSP) and GeoMx Whole Transcriptome Atlas (WTA) tools do use barcoded probes for capturing mRNA or proteins (Merritt et al., 2020; Roberts et al., 2021). However the GeoMx tools

require user input boundaries akin to microdissection, which leads me to classify it with those methods (see Section 2.2.4). Another method sometimes considered in this category is APEX-seq which depends on transgenic lines (see Section 2.2.4) (Fazal et al., 2019; Padrón, Iwasaki, and Ingolia, 2019).

2.2.2 *In-situ* hybridization

In situ hybridization is the longest existing spatially resolved method with the first application to mRNA in 1982 (Singer and Ward, 1982). Modern-day smISH usually targets multiple probes, conjugated with fluorophores that can be detected with microscopy, to the same transcript. In contrast to capture-based methods, ISH approaches are not whole-transcriptome and by design require unique probes for each gene (Table 2.1). In this manner, ISH approaches have suffered from low-throughput in terms of number of genes sampled per tissue sample. Another drawback of ISH methods is that they notoriously have a low signal-to-noise ratio due to high levels of tissue autofluorescence during imaging and high levels of non-specific probe binding. Further diminishing signal-to-noise ratio is overcrowding of detected molecules in the tissue which can make neighboring signals hard to resolve. Combating these issues, recently, there has been an advent of new ISH technologies seeking to multiplex and increase signal through (1) multiple rounds of ISH in the same tissue with probe stripping in between, (2) by designing probes so that multiple genes can be assayed at once, or (3) some combination of the two (Asp, Bergensträhle, and Lundeberg, 2020). Today, ISH methods have very high single molecule resolution and high sensitivity in probing target genes.

One of the first multiplexing efforts was sequential fluorescence *in situ* hybridization (seqFISH), which uses multiple rounds of hybridization to the same transcript to increase the signal-to-noise ratio (Lubeck et al., 2014; Shah et al., 2016). This is a labor and time intensive process. Improving on seqFISH is multiplexed error robust FISH (MERFISH), which saves time by hybridizing non-readout probes to a transcript and then using subsequent rounds of fluorophore-conjugated probes hybridizing to the initial probes for readout (Chen et al., 2015). The first MERFISH publication reported detecting 1,000

different mRNAs (Chen et al., 2015), while subsequent MERFISH publications have increased the number of different genes detected up to ~10,000 (Xia et al., 2019). Implementing a synonymous approach, seqFISH+ was introduced, reporting ~10,000 unique genes detected (Eng et al., 2019). Most recently a multi-omic approach combining RNA seqFISH+ with protein antibody immunofluorescence and DNA seqFISH+, to assay chromatin interactions and histone modifications, was reported (Takei et al., 2021a; Takei et al., 2021b).

Tangentially, ouroboros smFISH (osmFISH) was developed by hybridizing read-out probes to different mRNAs over multiple rounds of hybridization inter-layered with probe stripping (Codeluppi et al., 2018). In theory, such an approach is limitless in terms of genes targeted, but in practice the tissue can only withstand so many rounds of stripping.

As far as commercialization of ISH methods, the current readily available approach is called RNA scope (Wang et al., 2012). RNA scope uses a two step hybridization process with primary probes targeting transcripts and secondary probes binding to primary probes for read-out. RNA scope increases the signal-to-noise ratio, here the ratio of correctly bound ISH probe signal relative to background and/or off-target binding. This increase in signal-to-noise is accomplished by building 'trees' of multiple secondary probes binding to the primary probes. RNA scope multiplexing is currently limited to tens of genes. Additionally, MERFISH is currently under-going the process of commercialization through spin-out biotech start-up Vizgen.

As one of the oldest spatially resolved methods, ISH methods are well established across biological fields with optimized protocols for many types of tissue. In addition, most newly developed methods are compared to smISH as a sort of gold standard. This is particularly true of murine-based neuroscience where a publicly available central resource was developed by the Allen Institute of whole-transcriptome smISH on the entire adult mouse brain (see Section 2.4.1) (Lein et al., 2007). While generally lower-throughput and hard to design libraries for new species, ISH methods are ideal for smaller, targeted experiments. ISH methods will always serve as a validation for other whole-transcriptome and targeted approaches.

2.2.3 *In situ* sequencing

Similar to ISH, most *in situ* sequencing methods are targeted in terms of genes detected and use imaging as a read-out (Table 2.1). In ISS approaches, sequencing of transcripts or hybridized probes is done directly in the tissue themselves. ISS was first demonstrated using padlock probes in 2013 (Ke et al., 2013). The general idea is that mRNA is reverse transcribed into cDNA, which can then be bound by DNA padlock probes. Briefly, padlock probes are probes whose ends are designed to hybridize adjacently (sometimes with a gap) with the target molecule. The padlock probes take on a circular shape after hybridization. Here, the probes are ligated, amplified using rolling circle amplification (RCA), then sequenced. Within this general protocol, this paper demonstrated using padlock probes with and without a gap between the hybridizing regions of the padlock probe. When probes contain a gap, it is then filled by DNA polymerization, complementary to the bound cDNA. Though readout in either case is limited to 4 basepairs (bp), the gap-filling allows for the detection of single nucleotide variants. A draw back to ISS is that RCA takes a lot of physical space which limits the number of transcripts that can be detected. Additionally, the short 4bp readout length also limits the number of unique barcodes and, thus, targets. Short sequence length places a lower upper limit on unique sequences relative to longer lengths.

Iterating on this idea, barcode *in situ* targeted sequencing, or BaristaSeq, built on the gap-filling padlock probe ISS approach (Chen et al., 2017). Through multiple fronts of optimization, BaristaSeq is able to achieve sequencing of up to 15bp length. Notably, the polymerase used for gap-filling was switched and an extra cross-linking step was added prior to amplification to stabilize the amplicons. Additionally, sequencing chemistry was switched to Illumina sequencing-by-synthesis as opposed to sequencing-by-ligation as in the prior method; this may have helped increase signal. The increased gap-filling length of BaristaSeq is critical to assay more transcript diversity and was later modified to detect neuronal projections using a barcoding strategy (see Section 2.4.3) (Chen et al., 2019).

Most recently, Spatially resolved Transcript Amplicon Readout Mapping, STARmap,

was introduced which builds on these padlock ISS methods by removing the reverse transcription step (Wang et al., 2018). This was accomplished by hybridizing a second probe next to the initial padlock probe on the same transcript, which they call a SNAIL probe (based on its schematic appearance, or as an acronym for specific amplification of nucleic acids via intramolecular ligation). When both probes are hybridized to the same transcript, the padlock probe can circularize for RCA. STARmap, notably, reports 3-dimensional spatial resolution; after amplification, the nanoball amplicons are embedded in a hydrogel and the tissue is cleared prior to sequencing. This tissue clearing step allows molecules to be detected in 3-D space within a tissue. Similar to the initial ISS, STARmap sequences a 5bp barcode.

Tangentially, in 2015, fluorescent *in situ* sequencing of RNA, FISSEQ, was introduced (Lee et al., 2015). Contrary to all the above ISS methods, FISSEQ was revolutionary as a non-targeted approach. FISSEQ uses reverse transcription followed by RCA. FISSEQ takes advantage of a similar cross-linking step as Barista-seq (inspiring this approach in Barista-seq) to stabilize the amplicons for multiple rounds of sequencing. FISSEQ reports 30bp sequencing by ligation of 8000 genes. Since the DNA nanoballs created by RCA are large, FISSEQ can only sequence a random subset of detected transcripts at once. Recently, FISSEQ was improved upon by In Situ Transcriptome Accessibility Sequencing, or INSTA-seq (Fürth, Hatini, and Lee, 2019). INSTA-seq pairs FISSEQ-like ISS that sequences 5 to 6bp on each end of the transcript with regular next-generation sequencing of full reads. In this way INSTA-seq has both spatial location and full transcript reads, which allows access to information about variants.

ISS methods are generally very laborious and hard to execute. Many of these techniques were rarely demonstrated outside of the laboratories that invented them. Accessibility of ISS will likely only occur with commercialization. In 2016, a spin-out biotech start-up called ReadCoor was founded around FISSEQ. ReadCoor recently presented FISSEQ wrapped into a multi-omic platform. In addition Cartana was founded around the first padlock ISS approach. Both Cartana and ReadCoor were acquired by 10x Genomics, the company responsible for commercializing ST as Visium, in 2020. If made less-laborious

and more easily accessible, ISS methods provide an interesting high spatial-resolution alternative to capture-based methods with the added potential of detecting variants not possible with ISH methods.

2.2.4 Microdissection and other methods

In this section, I will discuss a few other techniques that do not fall neatly into the above three categories. The biggest remaining category of spatially resolved transcriptomics approaches is based on microdissection (Figure 2.2). Perhaps the most well known approach is simply laser capture microdissection (LCM) followed by RNA-seq of each of these regions separately (Emmert-Buck et al., 1996). This is a laborious process. (A similar approach was used for sequencing region-specific barcodes used to map neuronal projections (Huang et al., 2020).) Similarly, researchers have applied cryosectioning prior to RNA-seq to obtain spatial resolution (Junker et al., 2014; Asp, Bergenstråhle, and Lundeberg, 2020). Another way to 'microdissect' tissue prior to sequencing is through photoactivation. In this approach cells of interest are selected using photoactivation followed by cell sorting and sequencing (reviewed in Asp, Bergenstråhle, and Lundeberg, 2020). Methods of this class include TIVA and NICHE-seq (Lovatt et al., 2014; Medaglia et al., 2017). Similarly, ZipSeq uses photocaged oligonucleotide probes applied to a tissue that can then be activated using a microscope light to select and simultaneously image spatial regions of interest (ROIs) (Hu et al., 2020b).

Classified by other researchers as a capture-based method (Asp, Bergenstråhle, and Lundeberg, 2020), there are also the previously mentioned GeoMx DSP and WTA tools (Merritt et al., 2020; Roberts et al., 2021). While this approach applies probes to an entire tissue sampled, the ROIs must be manually selected by the user to cleave the barcoding probes from the RNA-bound region for subsequent read-out. GeoMx can also notably be applied to obtain spatial proteomics data by using antibodies bound to photocleavable oligo barcodes. The ROI selection step here appears more similar to 'microdissection' followed by read-out than the other capture-based methods.

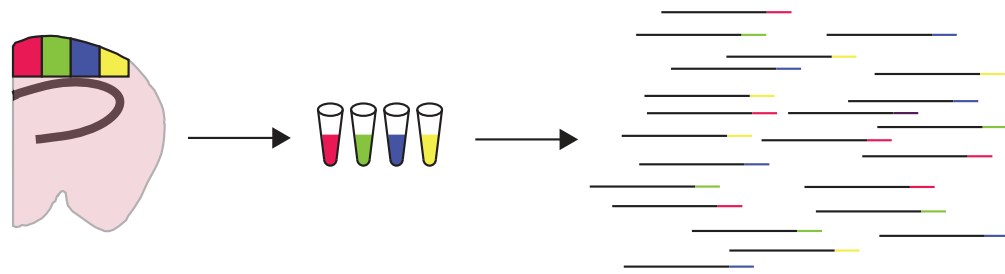


FIGURE 2.2: **Simplified schematic of microdissection and other manual tissue selection spatially-resolved transcriptomics methods.** Left: Example tissue with schematic micro-dissected regions depicted in different colors. Middle: Extraction of mRNA from each of the regions into separate tubes for library preparation. Right: After sequencing, mRNA reads include an index region that indicates which sample (micro-dissected region) it came from.

Another conceptually straightforward way to obtain spatially resolved gene expression is to take advantage of cell-type or region-specific transgenic lines. By design, this approach requires species-, cell-type-, and region-specific optimization: the creation of transgenic lines. This of, course, is limited only to model organisms that are genetically well-characterized. One example of such an approach is APEX-seq. Here, cell lines expressing APEX2 in specific regions of a cell can tag spatially local mRNAs that can then be isolated for sequencing (Fazal et al., 2019; Padrón, Iwasaki, and Ingolia, 2019).

As previously mentioned, imaging-based approaches such as ISH and especially ISS can suffer from over-crowding diminishing the signal-to-noise ratio. One way to circumvent this and provide access to more transcripts is by combining many of the above techniques with expansion microscopy. Expansion microscopy essentially embeds tissue in a highly optimized polyelectrolyte gel that expands with the addition of water (Chen, Tillberg, and Boyden, 2015). The chemistry of this technology is similar to how diapers work (Lee Henry, personal communication). One technology that has been combined with expansion microscopy is MERFISH (Wang, Moffitt, and Zhuang, 2018). Additionally, the same research group responsible for expansion microscopy, also introduced expansion sequencing, or ExSeq (Alon et al., 2021).

A final loose classification of methods are those that take advantage of mathematical tricks for spatially resolved expression. One example is DNA microscopy which reports spatial positioning of molecules relative to other molecules (Weinstein, Regev, and Zhang, 2019). cDNA molecules are randomly tagged with unique barcodes and amplified with overlap extension primers. The tagged molecules can then hybridize with other tagged molecules thanks to the extended ends. These concatenated molecules are read out by sequencing and the rate of concatenation two molecules reflects the distance between the two originally tagged molecules. A second example is CISI, or composite *in situ* imaging (Cleary et al., 2021). CISI limits hybridization imaging cycles by using composite probes of up to 10 genes per probe. These probes are designed based on simulations to include a subset of genes (~40) that are representative of gene modules identified using scRNA-seq data. After imaging and segmentation, the ISH data must then be decompressed. These methods are intellectually very interesting, but unlikely to have much practical value in other applications outside of the initial proof-of-principle.

2.2.5 Trade-offs of the different spatially resolved transcriptomics approaches and concluding thoughts

While the dividing lines between the methods classified above are permissive, the trade-offs of these different spatially resolved transcriptomics techniques are real. Traditionally ISH and ISS methods offer the highest spatial resolution, but are limited in the number of transcripts that can be assayed at once. With additional molecules tagged, the ability to resolve neighboring fluorophore signals drops. This is sometimes remedied with multiple rounds of hybridization or sequencing, but there is limit on the number of rounds a tissue can withstand while remaining intact and without drift in the field of view for microscopy.

Contra ISH, some ISS and capture-based methods have the ability to detect variants. ISH read-out does not involve any sort of sequencing or polymerization of complementary oligos; the signal comes from a pre-designed ISH probe recognizing a complementary mRNA transcript. Some ISS and capture-based methods have the ability to

sequence partial or whole transcripts which can allow for the detection of single nucleotide variants or splice isoforms. Further contra ISH, some ISS (e.g. FISSEQ) and all capture-based methods detect the whole transcriptome without the need to design gene specific probes. This is extremely powerful for discovery-based science where all mRNAs can be unbiasedly assayed without prior information.

Traditionally, capture-based methods have reported lower spatial resolution compared with ISH and ISS, but with each new capture-based method (e.g. seq-scope) the resolution has become increasingly on par. Combined with the high-throughput in gene space, capture-based methods are poised to overtake other spatially resolved approaches. However, increased ease of use and accessibility of these highly-technical approaches through automation and/or commercialization is required. Some methods already report some automation (Vickovic et al., 2020; Codeluppi et al., 2018) and other efforts are underway (as discussed throughout above on commercialization). Another important caveat, is that the increased spatial resolution of capture-based methods does not always track with an increased capture efficiency of transcripts. High spatial and transcript resolution is needed to truly supplant current single-cell sequencing methods that lack spatial resolution.

Focusing in on ISH and ST, the two tools used to collect the datasets used in Chapter 3, reveals some real differences between the two methods. ISH, as previously described, is an imaging based method. For highly expressed genes, there is a maximum saturation of microscopy detection for highly crowded signals that once reached cannot be increased (Levsky and Singer, 2003). This could mean that genes that are very highly expressed are not resolved and thus do not appear quantitatively different from genes that are slightly less highly expressed. In contrast, ST is a sequencing based method. There is no theoretical upper bound in detection of highly expressed genes. However, non-linear amplification of RNA in preparation for sequencing can lead to larger differences between highly and lowly expressed genes. Relative, not absolute, expression is more consistent across sequencing assays (Su et al., 2014). On the other hand, for lowly expressed genes, the high sensitivity specific gene probes in ISH are more likely to capture expression of these genes compared to ST where a finite number of generic probes could miss these

genes that are already present in low numbers. In ST, the absence or low numbers of these lowly expressed genes can be further exaggerated in amplification for sequencing. In these ways, there is a difference in dynamic range of expression detected by these two methods.

Multi-omic approaches (e.g. DBiT-seq, DNaseqFISH+, etc.) have the potential to offer a ground truth for computational multi-omic integration. These methods combine the collection of spatially-resolved multi-omic information from the same physical cells or tissues. Judging from the publications, multi-omic approaches involving proteomics tends to still be quite laborious. Regardless having multi-omic data from one source can act as a sort of decryption key for computational methods that integrate across various sources and types of omics data (see Section 2.3).

Notably, while many spatial methods are single-cell or lower in resolution, they often require joint analysis with scRNA-seq data for cell-type specific spatial patterning (e.g. HDST, slide-seq V1, etc.). Further, holding aside issues of tissue dissociation and doublets, single-cell methods are truly single-cell, while spatial methods depend on (often) probabilistic cell segmentation and transcript assignment. (One exception to this is sci-Space, who's addition of cell indexing after spatial barcoding renders its single-cell resolution similar to that of single-cell methods.) Further, high-throughput and high cell count scRNA-seq is already readily available at reasonable cost. Whether these 'drawbacks' of spatial compared to single-cell will hold, only time will tell. At this writing, spatial approaches are already used in complement with other RNA-seq approaches, but are well-poised to in time perhaps displace the need for current scRNA-seq approaches.

2.3 Spatial techniques, computational

Following the explosion of experimental spatially resolved transcriptomics methods, there has been an advent of new computational analysis tools for this class of data. Computational tools for spatial analysis usually pre-process spatial data, define spatially differentially expressed (DE) or marker genes, integrate spatial data with scRNA-seq or bulk RNA-seq data, or deconvolve non-cellular spatial data to single-cell resolution, to name a few examples (Table 2.4). Most currently available tools perform one or multiple

TABLE 2.4: Examples of computational spatially resolved transcriptomics tools.

Tool-type	Examples
Pre-processing and spatial analysis frameworks	starfish, Squidpy, SpatialExperiment, Giotto
Experimental tool specific pre-processing frameworks	STpipeline, Space Ranger, SMART-Q, dotdotdot
Identification of spatially DE genes	spatialDE, trendsceek, SPARK, spaGCN, stlearn
Integration with scRNA-seq or bulk	Seurat v3, Tangram
Deconvolution	cell2location, Giotto, Tangram
Early tools mapping spatial information to single-cell data	DistMap, NovoSpaRc

of these tasks. Notably, many of these techniques were published only in the last few years and it seems weekly that many new ones are popping up on bioRxiv. In this section, I will explore a few examples of techniques performing tasks in each of these categories. The focus of this section is on the general classes of tasks performed by spatial computational methods illustrated with examples; with exceptions, I will generally not include as examples the analyses done in experimental publications (see Section 2.2). Incorporating these elements, other researchers have provided more thorough, tour-de-force reviews (Moses and Pachter, 2021; Longo et al., 2021). Departing from the organization of the previous section, many of these tools often perform more than one of the above outlined tasks and will show up in more than one subsection below.

2.3.1 Pre-processing of spatial data

As there are a diversity of spatially resolved transcriptomics approaches, there is an appropriate diversity in analysis tools. Most spatial methods have a final read-out of either regular next-generation sequencing or imaging (see Section 2.2). (While the core of next-generation sequencing is also imaging, the image processing required there is well standardized with Illumina’s Real-Time Analysis software, and similar software for other sequencing platforms, during sequencing so that the output is flat text files of sequenced-reads and not images.) So, in general, pre-processing of spatial data looks very different

depending on whether the read-out is sequencing- or imaging-based.

For sequencing based spatial methods, older bioinformatics tools for bulk RNA-seq and scRNA-seq are easily re-purposed for initial spatial data processing. For example for read alignment, Space Ranger from 10x Genomics uses a wrapper around STAR for their Visium platform as STpipeline did earlier for ST (Navarro et al., 2017). For image read-out tools such as most ISS and ISH methods, pre-processing usually borrows from the image analysis toolkit. Tasks such as cell segmentation, barcode readout (turning cycles of fluorescent images into sequence reads), and assigning reads to transcripts are common. These tasks can and are accomplished through standard programming languages such as R, python, and MATLAB (reviewed in Moses and Pachter, 2021). In addition, some researchers use canonical microscopy analysis tools such as imageJ/FIJI (Chen et al., 2017). For a spatial-focused re-implementation of some older image processing tools, such as segmentation of an image to identify cells or registration of an image to a common reference, the Starfish toolkit was developed by the Chan-Zuckerberg Institute (CZI) (Perkel, 2019; Ganguli, Carr, and Long, 2018). Similarly, SMART-Q (abbreviation for Single-Molecule Automatic RNA Transcription Quantification) iterates on starfish, with more modular design and optimization for application to RNAscope data (Yang et al., 2020). Implemented in MATLAB, unlike starfish and SMART-Q in python, dotdotdot is a toolkit also used for processing ISH data, including RNAscope, containing some additional statistical analysis tools (Maynard et al., 2020). As shown by these examples, there are a variety of tools for analyzing imaging- and sequencing-based spatial data spanning many programming languages, software design, and statistical approaches.

2.3.2 Assay-independent spatial analysis frameworks

In the last two years, there have been a variety of standardized frameworks pre-printed or published that are more comprehensive, experimental platform-independent computational analysis platforms. These toolkits usually contain a variety of exploratory data analysis and visualization tasks applied to a standardized way of storing spatial data. Often, these frameworks are compatible with other spatial analysis tools either by easily in-taking

their output or by pre-processing and outputting data in a format required by other tools. Some of these comprehensive toolkits have evolved out of single cell packages such as Squidpy and SpatialExperiment (Palla et al., 2021; Righelli et al., 2021), while others are original to spatial analysis such as Giotto (Dries et al., 2021).

Squidpy, or Spatial Quantification of Molecular Data in Python, is essentially a wrapper bridging the world of molecular omics analysis and image analysis for the benefit of spatial transcriptomics analysis (Palla et al., 2021). Squidpy is built on Scanpy (Wolf, Angerer, and Theis, 2018) and other python libraries such as the image analysis library scikit-image (Walt et al., 2014). The core of Squidpy is the storing of data as a neighborhood graph in addition to the acquired images. The neighborhood graph allows Squidpy to be agnostic to the spatial technology used (see Section 2.2), just by defining the nodes according to the spatial resolution of the assay used. In this framework, many statistical analyses can be done such as using calculating Moran's I to identify spatial variation in expression. Further, pre-processing steps such as segmentation and feature extraction can be done in Squidpy and passed into other tools like Tangram and Cell2Location (discussed in Section 2.3.4 and 2.2.5 below) which require cell-segmented data (Biancalani et al., 2020; Kleshchevnikov et al., 2020). Squidpy provides a well-defined framework for spatial data analysis using bioinformatics and modern machine learning in python together.

Similar to Squidpy, in the R data analysis world, there is SpatialExperiment (Righelli et al., 2021). SpatialExperiment expands on its precursor for single cell data, SingleCellExperiment (Amezquita et al., 2020). In addition to the capabilities of SingleCellExperiment, SpatialExperiment stores spatial coordinates of data and, optionally, the associated imaging data. SpatialExperiment provides a consistent framework for R users to analyze spatial data through exploratory analysis and visualization. Further, SpatialExperiment also allows users to directly import the output of 10x Genomics' Space Ranger pipeline. Notably, SingleCellExperiment and SpatialExperiment are strong frameworks to store and work with their respective data types because of their consistency in storing metadata. Like Squidpy, SpatialExperiment also interfaces with other spatial tools to easily perform analyses implemented in other packages.

A toolkit original to spatial analysis is the Giotto package implemented in R (Dries et al., 2021). It was one of the first available spatial analysis toolkits that is not specific to any one spatial experimental technique. Giotto allows for analyzing and visualizing spatial data such as finding spatial patterns of expression and spatial cell-type enrichment. For analyzing capture-based approaches, Giotto integrates with scRNA-seq data (more on this in Section 2.3.4). These comprehensive analysis and visualization frameworks provide an easy entrance into analyzing spatially resolved transcriptomic data for a variety of users.

2.3.3 Identification of spatially differentially expressed or spatial marker genes

After data pre-processing, some basic analysis can then be done on spatially resolved transcriptomics data. One common first question is the identification of spatial patterning of expression through the identification of spatially differentially expressed (DE) genes or spatial marker genes. One such method is SpatialDE (Svensson, Teichmann, and Stegle, 2018). SpatialDE uses Gaussian process regression to find spatial differences in expression by separating spatial variance from the non-spatial ones. SpatialDE can also cluster spatially based on expression to generate what the authors term a "automatic expression histology" (Svensson, Teichmann, and Stegle, 2018). A second method is trendsceek which uses a marked point process to implement a non-parametric approach (Edsgård, Johnsson, and Sandberg, 2018). Trendsceek identifies the distribution of spatial locations of each cell and models it as a joint distribution with gene expression. Then in a pairwise manner, it determines if there is a significant relationship between the cells and their expression given the distance of the two points. Spatial patterning of expression is significant for the two cells if it is different from the null random distribution of the associated expression with cells. Deviating from both of these approaches is SPARK, or spatial pattern recognition via kernels (Sun, Zhu, and Zhou, 2020). SPARK works directly with un-normalized expression data to identify spatial differentially expressed genes using generalized linear modeling. This is in contrast to SpatialDE and trendsceek, which uses normalized data. Sometimes, the choice of normalization can alter the dynamic range of the original data. The authors claim that without normalizing, SPARK can perform better powered analysis.

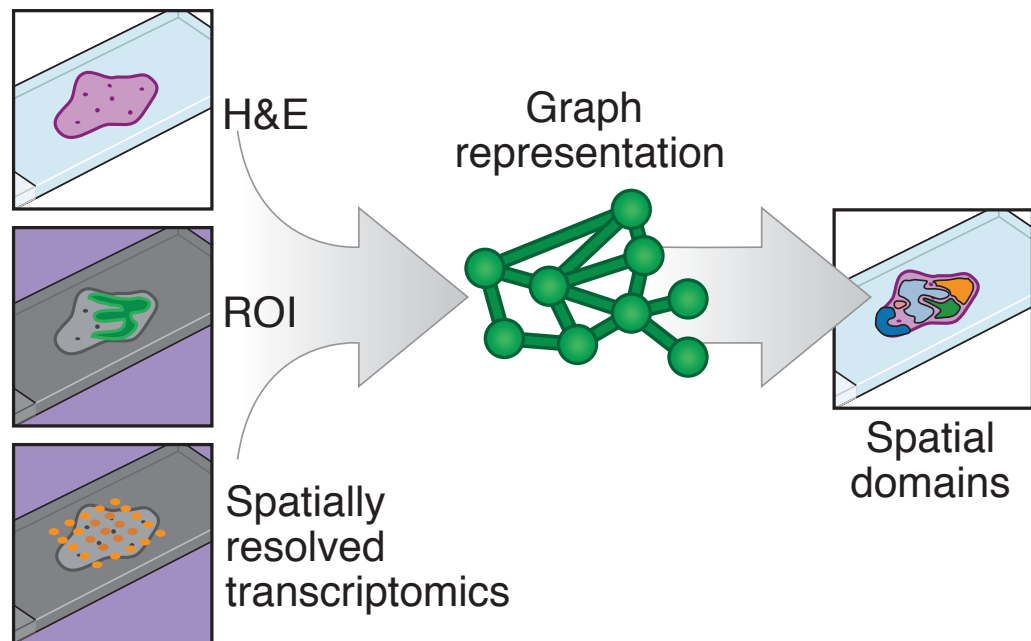


FIGURE 2.3: **Schematic of SpaGCN spatial analysis tool.** SpaGCN integrates histological information, user-defined region of interest (ROI), and spatial transcriptomics into a graph convolutional network (GCN) and performs unsupervised clustering on the graph representation to arrive at a set of spatial domains. *This figure was created by Daniel Fürth.*

Following this initial set of tools, recently there were a couple of methods shared that insert a spatial clustering step prior to the identification of spatially DE genes. SpaGCN, or spatial graph convolutional network, is a flexible spatial transcriptomics analysis tool that combines expression, location, and histology (Figure 2.3) (Hu et al., 2020a). This tool uses a graph convolutional network to identify spatial boundaries, or clusters, from the data paired with subsequent identification of differentially expressed genes between the spatial clusters, which they refer to as spatially variable genes (SVGs). When singular genes cannot be found to distinguish a cluster, a metagene combination of multiple genes is proposed as a cluster marker. Briefly, a graph convolution network is just a specific architecture for neural networks that includes a convolution step(s) where information is shared across neighboring parts of the network (like convolutional neural networks) and can work with more randomly structured data (unlike convolutional neural networks). SpaGCN has been demonstrated on a variety of spatial technologies including ST/10x Visium and MERFISH (see Section 2.2).

A similar tool, also incorporating histology is *stlearn* (Pham et al., 2020). Similar to *spaGCN*, *stlearn* proposes a spatial clustering of spatially-resolved transcriptomic data by incorporating information from histology. To extract information from the histology images, *stlearn* uses a pre-trained convolutional neural network (trained in ImageNet, not biological images necessarily) to extract numerical features, while *spaGCN* uses a simple normalized pixel intensity. *Stlearn* then uses this alongside expression and location information to cluster with Louvain or K-means. After proposing spatial clusters, *stlearn* goes on to include a sort of spatial pseudotime and cell-cell interaction analysis components. *Stlearn* is demonstrated only on ST and 10xVisium so far.

Worth mentioning are other emerging tools such as *Tangram* and *Seurat* which also include the identification of spatial DE genes, but will be discussed in following sections since their innovations are better categorized in their respective sections (Biancalani et al., 2020; Stuart et al., 2019). Stepping back, while the identification of spatial patterning is useful (e.g. capturing brain morphology with expression or identifying tumor heterogeneity), it is important to keep in mind that the particular proposed clustering may not be representative of the biological process that generated it.

2.3.4 Integration of spatial data with single-cell RNA-seq

As detailed in Section 2.2.5, traditionally there is a trade-off between the different types of spatially resolved transcriptomics experiments. High spatial resolution methods such as ISH and ISS tend to only be able to sample a subset of the transcriptome, while whole-transcriptome capture-based methods tend to have lower spatial resolution. While new experimental methods are rapidly closing this gap, until these tools are readily available, computational integration between datasets can bridge these two regimes. A popular approach is to integrate scRNA-seq with spatial methods, particularly with ISH and ISS approaches.

Here, we highlight a few examples of data integration between single-cell and spatial integration methods. An early integration method, published in 2015, used a straightforward approach to match expression profiles across ISH and scRNA-seq (Satija et al.,

2015). The ISH data was used to determine whether genes were expressed or not on a binary scale across defined spatial bins to which the single-cell data was mapped. Using co-expression the expression of other ISH genes could then be imputed in the single-cell data. This method was applied to the zebrafish embryo, which as a well-studied model organism likely made this approach possible. A later paper from the same group demonstrated a second method of spatial and single-cell data integration as part of a larger toolkit with integration of many data types (i.e. scRNA-seq and scATAC-seq) (Stuart et al., 2019). This new approach relied on the identification of so-called "anchors," or cells with pairwise correlation across datasets. This integration allowed ISH data to be imputed with more genes from the single cell data and cell types to be transferred from single-cell to partial data. It was demonstrated on combining scRNA-seq (SMARTseq2) with the ISH method STARmap. These methods were incorporated as part of Seurat v3.

Another recent tool, Tangram also integrates single-cell or single-nuclei data with spatial methods (Figure 2.4) (Biancalani et al., 2020). The basic idea is to randomly assign the sequenced cells or nuclei in space then compute an objection function to maximize both the similarity between cell density and gene expression. The cells are then rearranged in space to maximize the correlation. Tangram is compatible with most published spatial methods at the time of its publishing. Following integration, a number of analysis tasks can be accomplished using Tangram, such as imputing additional genes in spatial data that is not transcriptome-wide or deconvolving spatial data that is not cellular-resolution into cell type proportions.

While many spatial analysis tools perform data integration between single-cell and spatial methods, many of these tools are to achieve another purpose beyond integration. For instance NovoSpaRc and DistMap use integration to identify spatial patterning of single-cell data. These early approaches are less about the integration itself and rather about assigning spatial location to single-cell data (Nitzan et al., 2019; Karaikos et al., 2017) (see Section 2.3.6). Other tools, cell2location, Giotto, and Tangram integrate for the purpose of deconvolving non-cellular spatial data and assigning cell types, as discussed in the following section (Kleshchevnikov et al., 2020; Dries et al., 2021) (see Section 2.3.5).

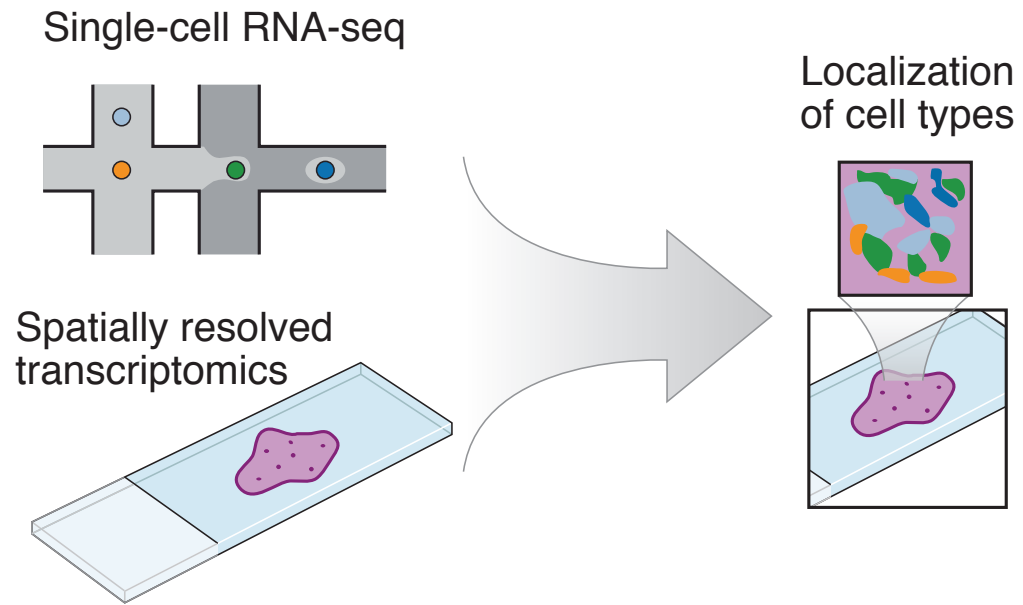


FIGURE 2.4: **Schematic of Tangram spatial analysis tool.** Tangram aligns single-cell data with spatially resolved data to arrive at imputed and deconvolved spatial domains with single-cell like qualities. *This figure was created by Daniel Fürth.*

2.3.5 Deconvolution of non-single-cell resolution spatial data

Following from integration described above, a second purpose of integrating single-cell and spatial data is to deconvolve non-cellular resolution spatial data. Many of the original capture-based methods, in particular, barcode spots that can map to multiple cells (see Section 2.2.1). Even some of the newer, higher resolution methods could get transcripts from multiple cells if the spots happen to line up with a cell boundary.

One such deconvolution tool is Cell2Location (Kleshchevnikov et al., 2020). In its publication, Cell2Location was tested on both Slide-seq and 10x Visium data across different type of tissues. Cell2Location works in a two step process where it first estimates cell type expression profiles from the single cell data by clustering the cells then averaging the expression profile for each cluster. In the second step, these expression profiles are used to decompose the expression data from the sampled spatial spots by modeling the spatial expression as a linear regression of the cell type profiles.

Giotto, as discussed above, also has a data integration component (Dries et al.,

2021) (see Section 2.3.2). For lower spatial resolution capture data, additional gene expression information in the form of either marker gene lists or single-cell data can be used to estimate the relative proportions of cell types in a given sampled spot. This is done by calculating an enrichment score between the markers and the fold change or ranking of those genes in the spatial data spot. Note, methods for deconvolving spatial data also exist in Tangram, Seurat, and elsewhere (Biancalani et al., 2020; Stuart et al., 2019; Moses and Pachter, 2021). Deconvolution to identify single cells or cell type proportions is particularly useful for whole-transcriptome, capture-based approaches that lack cellular resolution.

2.3.6 Mapping spatial information onto single-cell data

Stepping back, early tools combining scRNA-seq with spatially-resolved data were more focused on mapping single-cell data in space than on true harmonization of the datasets through integration. Though the goal of finding spatially-resolved expression is the same, here I describe these early tools in a separate sub-section since they are distinct from more recent integration tools.

Published in 2017, DistMap was used to reconstruct the *Drosophila* embryo from single-cell data (Karaïskos et al., 2017). Essentially, similar transcriptomes between the single-cells were clustered to get a spatial mapping. A reference spatial dataset in *Drosophila* was spatially binned and expression of genes in each of these bins was binarized. The single-cells then got a distributed score based on their mapping to each of these bins. For all pairwise comparisons of each cell to each bin, the number of genes that were on or off in both the bin and cell was compared to the number of genes that disagreed between the bin and cell. Notably, the success of DistMap depended on the availability of the central spatially-resolved resource for *Drosophila* embryos. Today, with readily available whole-transcriptome tools, the integration approach proposed in DistMap could likely be modified to be applicable more generally.

Later in 2019, NovoSpaRc another tool mapping single-cells in space was introduced (Nitzan et al., 2019). NovoSpaRc can actually function with or without spatial data.

In essence, NovoSpaRc simply arranges sequenced single-cells in space based on similarities and differences in their expression patterns. When using spatial data, a distance matrix of the cells in expression space and physical space is calculated and minimizing this distance is framed as an optimization problem. The goal of these tools is simply to find the spatial distribution of single-cell data and not necessarily to integrate with spatial data.

2.3.7 Conclusions

Some of the text in this section previously appeared in the Nature Methods News & Views piece titled "Integrative analysis methods to bridge trade-offs in spatial transcriptomics data," which was authored jointly by Shaina Lu, Daniel Fürth, and Jesse Gillis. The full published text is available in Appendix B. I wrote the original text and contributed to subsequent rounds of substantial editing. All text that appears in this section was authored by me.

To serve the rapidly evolving field of spatially-resolved transcriptomics, there are a variety of analytical tools available. In this non-comprehensive section alone, we have covered tens of tools all recently published within the last 5 years. Many of these tools are actually quite complementary to one another as they span different types of analysis for spatial data. They range from imputation of missing genes or resolution (deconvolution) to comprehensive frameworks for spatial bioinformatics. Beyond the types of tools covered here, there are additional methods that explore cell-cell interaction or incorporate the idea of pseudotime from single-cell analysis with spatial data (Pham et al., 2020). It is also worth noting that we did not cover analysis tools that are meant to resolve crowded ISH or ISS data during image processing such as BarDensr (Chen et al., 2021b) (see Section 2.2.3 and Section 2.3.5).

Stepping back, it is worth considering what spatial patterning of expression maps to biologically. By taking an agnostic, data-first approach, spatially-resolved transcriptomics analysis tools posit that the proposed particular clustering and subsequent identification of gene expression patterning could represent a biological process. While these tools can capture biological phenomenon such morphological patterns in the brain, one

proposed clustering does not necessarily capture the biological mechanism by which that dataset was generated. This is an important distinction to keep in mind.

On a more technical level, it is worth noting that the comparison of spatially-resolved transcriptomics analysis methods often depends on assessments that are quite qualitative in nature. In other words, spatial clustering methods or identification of spatial distributions of cell types, for example, are often visualized with microscopy images and said to be good representations when these computationally defined features match with cytoarchitecture and morphology of the tissue. There are some popular statistical measures (e.g. Moran's I), but these do not capture the performance of all classes of spatial analysis tasks. With so many new experimental methods developed for spatially-resolved transcriptomics in the last 5 years (Asp, Bergenstr hle, and Lundeberg, 2020; Moses and Pachter, 2021) (see Section 2.2), a proliferation of computational methods to analyze these assays reliably followed (Moses and Pachter, 2021; Longo et al., 2021). As with any new field, to better understand the pros and cons of the many spatial analysis tools, an independent, rigorous, and quantitative benchmarking across spatially resolved transcriptomics analysis tools is needed (Boulesteix, Lauer, and Eugster, 2013).

As experimental technologies continue to improve, the gap between high spatial resolution and percentage of the transcriptome assayed continues to dwindle. However, until new techniques that promise to be whole-transcriptome with sub-cellular resolution (Cho et al., 2021) are readily available and accessible, computational data integration is necessary to bridge this gap. Data integration for transcriptomics is an old field, reaching back to microarray data, with clustering and matching neighboring cells in expression space re-purposed here for spatial data. The promise of data integration approaches is a truly multi-modality understanding of biology through creation of large, integrated datasets such as the Human Cell Atlas (Regev et al., 2017; Regev et al., 2018). Indeed, adapting old data integration techniques in new frameworks for new spatial data is a first step toward truly harmonized datasets.

Additionally, as discussed previously (see Section 2.2.5), other experimental advances represent multi-omic approaches (Takei et al., 2021a; Deng et al., 2021b; Deng et

al., 2021a) that have the potential to offer a ground truth for computational multi-omic integration. By combining the collection of spatially-resolved, multi-omic data from the same physical sample, these methods can be used to train computational methods to extrapolate from these examples in integrating larger datasets. Such integration would complement these multi-omic experimental methods which are still quite laborious.

When analyzing data using these tools, it is important to keep in mind the biological interpretation, or potentially lack thereof, of the results. While improvements in experimental technology may make some applications of data integration less useful, being able to integrate across existing, large datasets will be critical for the longer-term future. Finally, with so many tools available, it is important to critically assess their relative strengths and weaknesses. Spatially-resolved transcriptomics has the potential to be the revolution of this decade, much as single-cell was for the last; these analysis tools will help to realize their potential.

2.4 Spatial expression combined with other data modalities in neuroscience

With its stereotyped sub-structure and transcriptional heterogeneity, the mammalian brain is an ideal system to apply spatially-resolved transcriptomics methods. As a well-established model organism, combined with the availability of central references, the mouse brain in particular is a perfect first system. Indeed, most spatially-resolved transcriptomics tools show a proof-of-principle in the mouse brain, especially in the olfactory bulb (e.g. Ståhl et al., 2016). In converse, the study of the brain also stands to benefit from the advent of spatial tools. A multi-modality understanding of the brain, in particular, is possible with spatial tools as a link between molecular properties such as expression and mesoscale properties such as projection patterns and anatomy (Lein, Borm, and Linnarsson, 2017). So far, the focus in this chapter has been on spatially-resolved transcriptomics more generally. In this section, I will explore the application of these techniques in the context of multi-modal approaches to neuroscience. I will also include multi-modal approaches that use expression, even without an explicit spatial component, that are particularly striking.

2.4.1 Early spatial expression in neuroscience (Allen Brain Atlas)

Before the recent advent of high-throughput spatially-resolved transcriptomics techniques (see Section 2.2, spatially-resolved expression was already well-established in neuroscience through the availability of large, central resources, namely the databases of the Allen Institute. In 2007, the Allen Institute published a whole-transcriptome, spatially-resolved database for the whole adult mouse brain (Lein et al., 2007). The Allen Brain Atlas adult mouse *in situ* hybridization (ABA) dataset consists of a transcriptome-wide assay of expression in inbred WT mice using single molecule ISH (Lein et al., 2007). To achieve this, each gene is considered as independent image series that are subsequently reconstructed to three dimensions and registered to the reference brain atlas in interlayered steps (Ng et al., 2007). There are 26,078 series, or experiments, across both coronal and sagittal planes with 19,942 unique genes represented. These 3D registered reconstructions are then segmented to 200m^3 voxels with an associated brain area label. There are 159,326 voxels, with 62,529 mapping to the brain. Gene expression for each of the assayed genes was quantified in these voxels from the imaged data as energy values which is defined as the sum of expression pixel intensity divided by the sum of all pixels.

Since publishing, the ABA has become a central resource for modern neuroscience with over 3000 papers citing to it. An obvious subset of these papers are those that iterate in knowledge and development of the ABA. These include reviews (Fornito, Arnatkevi, and Fulcher, 2018), tools for working with the data (e.g. ways of projecting 3D data to 2D) (Hawrylycz et al., 2011; Grange, Hawrylycz, and Mitra, 2013), expression analysis papers on the dataset (Henry and Hohmann, 2012; Bohland et al., 2010; Ko et al., 2013; Tan, French, and Pavlidis, 2013; Grange et al., 2014; Cohen et al., 2017), and papers linking expression and connectivity (Wolf et al., 2011; French and Pavlidis, 2011; French, Tan, and Pavlidis, 2011; Fulcher and Fornito, 2016). A large chunk of papers citing the ABA, focus on specific brain areas or genes and use the ABA as a starting place for building hypotheses, a comparison for validation, or both. For example, many of the new spatially-resolved expression techniques always compare to the ABA for validation (e.g. (Lee et al., 2015; Ståhl et al., 2016; Sun et al., 2021)). The ABA is also often used as a way

of adding *in situ* knowledge to traditional RNA-sequencing based studies. Exhibiting the wide influence of the ABA, additional papers citing to it span fields of biology and as a non-representative non-random sample, include: the UCSC genome browser, papers about machine learning, Autism spectrum disorder research, fear learning research, and much more.

Over the years, the Allen Institute has expanded to include an adult human brain atlas, developing mouse and human atlases, mouse spinal cord atlas, and more atlases in the same fashion of spatially-resolved expression assayed using a variety of genomics approaches. Through other research initiatives, comprehensive spatially-resolved resources exist in other fields (prior to the recent development of spatially-resolved experimental tools) as well. These are the databases that generally enabled the research mapping single-cell data in space described in Section 2.3.6. Despite the laboriousness of prior spatially-resolved expression databases, these resources were clearly central to modern neuroscience, enabling lines of research not otherwise possible.

2.4.2 A few examples of multi-modal neuroscience studies

With exceptions, until the last five years or so, molecular neuroscience (i.e. gene expression, methylation, etc.) and computational neuroscience (i.e. physiology, behavior, etc.) were very distinct fields. Rarely would papers published in neuroscience span these categories. In 2018, however, the Svoboda Lab published a paper incorporating single-cell expression, projection/morphology tracing, physiology, and behavior (Economo et al., 2018). This paper would quickly become an iconic reference for multi-modality neuroscience. Essentially, the researchers performed scRNA-seq of the motor cortex and visual cortex, but then decided to focus on pyramidal tract (PT) neurons of the motor cortex. They then did retrograde labeling of the PT neurons for several sub-cortical regions and found that neurons projecting to the medulla and thalamus corresponded to separate expression clusters. These retrograde labeled cells from the thalamus and medulla were then sequenced to get cell-type specific transcription. Finding differentially expressed genes from the ABA, the researchers then followed up with a couple of genes using smFISH and

found two genes that could distinguish the two PT cell types. Finally, they correlated these two cell types with different behavioral roles using projection specific recordings obtained through retrograde channelrhodopsin in a delayed response task. In summary, Svoboda and colleagues found two PT cell types in the motor cortex with distinct projections, expression, and behavior related signals.

The following year, the Hantman group published a paper combining projections, expression, and physiology in the thalamus (Phillips et al., 2019). Here, the researchers again used retrograde labeling, injecting at 8 different projection targets to label neurons in the thalamus. Labeled neurons were then microdissected, sorted, and sequenced. Clustering and dimensionality reduction on the top 500 DE genes revealed 5 major transcriptional divisions, with 2 of these corresponding only to singular thalamic areas. This left 3 major transcriptomic groups that were spread across thalamic nuclei projecting to motor, somatosensory, and visual cortex. The three groups named primary, secondary, and tertiary mapped to functions (primary containing unprocessed information, tertiary contains cognitive signals), topographical locations, and an increasing difference in expression between the groups. The transcriptional groups were validated with whole-cell patch clamping showing differences in action potential curves of the three groups and morphology showing the primary group projecting mostly to middle layers where raw projections would go, secondary group mostly to layer 1, and tertiary group least to layer 1. Transcriptional differences were then validated with single-cell sequencing of thalamic neurons which further showed that similar marker genes were present in clusters across regions that the thalamic neurons project to. Single-cell sequencing further revealed a continuous axis of cell types rather than distinct types which was validated with cell type markers probed with ISH. In summary, Hantman and colleagues found distinct thalamic neurons defined transcriptionally, that also had distinct physiology and morphology. These transcriptional groups were repeated across projection systems.

These of course are not the only multi-modal studies in neuroscience, but they are particularly striking as examples showcasing the potential of combining molecular and computational neuroscience.

2.4.3 New experimental tools for multi-modal neuroscience

Alongside these discovery-based multi-modal papers, a variety of tools for collection of more than one modality of data in the brain have also been recently published. Combining projections and spatial transcriptomics, Barcoded Anatomy Resolved by Sequencing or BARseq, was introduced in 2019 (Chen et al., 2019). BARseq builds on the previously described BARISTAseq *in situ* sequencing method (see Section 2.2.3) (Chen et al., 2017) by combining it with barcoded labeling of neuronal projections named MAPseq (Multiplexed Analysis of Projections by Sequencing) (Kebschull et al., 2016). BARseq allows for *in situ* sequencing of gene expression and barcoded neurons at their somas combined with regular bulk RNA sequencing of their barcoded projections at target sites. BARseq was later improved (and called BARseq2) to assay more endogenous genes (Sun et al., 2021). Another method combining transcriptomics and expression is Connect-seq (Hanchate et al., 2020). Connect-seq uses a cross-synapse retrograde virus that will only cross if cre is present in the infected neuron. After infection, the tissue is dissociated and sorted for fluorescent, infected cells prior to scRNA-seq. Transcriptomic and connectivity information can then be mapped to each other in the neurons. Yet another manuscript goes one step further and combines epigenetic information with projections (Zhang et al., 2020). Here, retrograde labeling was combined with single nucleus DNA methylation sequencing on cortical neurons labeled according to their long distance projections.

An example technique combining expression and physiology is patch-seq (Qiu et al., 2012; Cadwell et al., 2016; Fuzik et al., 2016; Scala et al., 2020; Gouwens et al., 2020). Patch-seq combines whole-cell patch clamping with scRNA-seq, generally providing electrophysiological and transcriptional information together from the same cells. The first proof-of-principle in 2012, only demonstrated this on 5 cells and did not even include physiology (Qiu et al., 2012). Here, patching was used just to isolate single cells for sequencing. In 2016, a couple of true patch-seq papers were published including papers that also combined patch-seq with other modalities such as morphology (Cadwell et al., 2016; Fuzik et al., 2016). In one paper, 58 neocortical cells were studied with whole-cell patch clamp electrophysiology followed by aspiration of the cell for sequencing (Cadwell et al.,

2016). There were two known types of cells with distinct firing patterns that were also shown to have unique transcriptional profiles. Later in 2020, as part of the BRAIN Initiative, over 1300 cells from the primary motor cortex (MOp) were profiled using patch-seq (Scala et al., 2020). The researchers demonstrated that broad transcriptomic types have distinct morphological and physiological profiles, but the correlation is not one-to-one and profiles are continuous within the transcriptomic types. Another paper from 2020, profiled 3,700 GABAergic neurons from the primary visual cortex (VISp) (Gouwens et al., 2020). Though there was some correspondence, the researchers found that there was generally a large discrepancy between transcriptional and morphological types.

In this section, I have focused on examples of techniques that bridge molecular information with other types of neuronal data. However, it is worth noting that there are a variety of papers and techniques that combine different types of molecular information applied or developed for the brain. For example, DNA seqFISH+ (see Section 2.2.2) which combines expression, protein detection, chromatin topology, and functional chromatin marks was applied to the adult mouse cerebral cortex (Takei et al., 2021a). With exceptions, many multi-modal approaches to neuroscience bridge molecular and mesoscale properties, such as projections and cell physiology, but exploration of emergent properties of behavior is generally still distinct.

2.5 The need for benchmarking across spatial techniques

Some of the text in this section was previously used in the cover letter for publication of the manuscript detailed in Chapter 3. I wrote the original text with guidance and editing from Jesse Gillis.

To obtain the ultimate goal of multi-modal studies, there is a need to first understand robustness within one type of data. As with any new technique, spatial approaches must be benchmarked against each other. The spatially resolved transcriptomics approaches introduced in Section 2.2 are all assaying the same biological phenomenon. Theoretically, this means that in the same biological system, datasets produced using one of these technologies should replicate with a second produced using a second technology.

Even if replication is not on the one-to-one raw data scale (due to differences in technologies on resolution, genes-assayed, etc.), replicability should minimally occur in biological conclusions drawn from the spatial data.

To perform this sort of benchmarking, we need a good model system. In the following chapter (Chapter 3), we exploit the stereotyped sub-structure and transcriptomic heterogeneity of the mature mouse brain, to assess the replicability of spatial gene expression assays. Specifically, we benchmark Spatial Transcriptomics (see Section 2.2.1), the precursor to the commercialized 10x Visium platform, relative to the painstakingly generated *in situ* hybridization dataset from the Allen Institute (see Section 2.4.1). Instead of focusing on strict replicability of the data per se, we focus on the conclusions that data allows us to draw, asking if brain sub-areas could be similarly learned using gene expression across the two datasets. Using linear modeling in a supervised learning framework, we principally find that brain areas are classifiable using gene expression, but that there is a discrepancy in performance between the two datasets. We follow this up to determine that the Spatial Transcriptomics dataset generalizes better to the Allen Institute’s dataset than the reverse. We dig into possible explanations for these observations, pointing out relevant examples and likely explanations that do much to illuminate best practices for future use of these methods.

Batch effects have plagued genomics since before microarrays and spatial transcriptomics appears to be no exception. While our analyses of generalizability are across reference data sets, our findings are of much the same kind and importance as batch effects, suggesting the same need for replication, randomization, and control. This work lays important groundwork for quantifying biases across spatial transcriptomics approaches both as the first comprehensive, cross-platform characterization of spatial gene expression assays and the first comprehensive benchmarking of the Allen dataset with an independent dataset. This work is an extension of the increased focus on evaluating high-throughput expression methods using computational approaches (Skinnider, Squair, and Foster, 2019) applied specifically to the rapidly developing field of high-throughput spatially resolved transcriptomics.

Chapter 3

Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain

This chapter is adapted from the PLOS Biology Methods and Resources Paper titled "Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain," which was authored jointly by Shaina Lu, Cantin Ortiz, Daniel Fürth, Stephan Fischer, Konstantinos Meletis, Anthony Zador, and Jesse Gillis. The full published text and supplementary materials are available in Appendix C and Lu et al., 2021.

While there are a multitude of computational analysis papers pertaining to spatially-resolved transcriptomics (see Section 2.3), most of these focus on tools for processing and analyzing collected data. To our knowledge, so far none of these focus on the replicability of independent spatial datasets in a comprehensive manner. Here, we seek to do just that between two independent, spatially-resolved datasets across the whole-transcriptome using the adult mouse brain as an ideal model system.

3.1 Introduction

In the last five years, there has been an explosion of spatially resolved transcriptomics techniques that have made it possible to easily sequence whole transcriptomes while

retaining fine-scale spatial information (Ståhl et al., 2016; Rodriques et al., 2019; Vickovic et al., 2019; Stickels et al., 2020; Asp, Bergenstråhle, and Lundeberg, 2020) (see Section 2.2). These new technologies are poised to be transformative across biology (Marx, 2021). Despite the recent proliferation and improvement of single-cell technologies, these technologies largely depend on tissue dissociation and thus lack information on the spatial origin of sequenced cells. New spatial sequencing tools fill this gap, allowing us to understand the spatial patterning of cell-type specific expression. The stereotyped spatial organization and transcriptional heterogeneity of the brain make it an especially appealing application of these new technologies. Spatial gene expression has the potential to serve as a link between the molecular, meso-scale, and emergent properties of the brain such as gene expression, circuitry, and behavior, respectively (Lein, Borm, and Linnarsson, 2017; Close, Long, and Zeng, 2021). This, in turn, could lead to tackling longstanding questions about the brain, such as how gene expression relates to connectivity of neurons or how spatial patterning of expression drives development. Emerging experimental approaches (Economio et al., 2018; Bendesky et al., 2017; Moffitt et al., 2018) and techniques (Cadwell et al., 2016; Hanchate et al., 2020; Chen et al., 2019; Huang et al., 2020; Sun et al., 2021) have already begun to link multi-source information from the mouse brain. However, in order to perform robust multi-modality studies, we must first assess replicability within one type of data. Given the potential of spatial transcriptomics approaches in neuroscience, the early availability of spatial data, and the stereotyped sub-structure, we use the adult mouse brain as a model system for a cross-platform characterization of spatial data.

Over a decade ago, the first whole-transcriptome, spatially-resolved gene expression dataset from the adult mouse brain was collected by the Allen Institute using *in situ* hybridization (ABA) (Lein et al., 2007; Ng et al., 2007). Since its release, this dataset has become a cornerstone for modern neurobiologists who often use it as a first point of reference for gene expression in the mouse brain. The generation of this dataset was a laborious effort requiring many years, the work of many scientists, and many sacrificed mice. The influx of technologies preserving the spatial origin of transcripts presents the opportunity to assess the generalizability of the ABA data for the first time. As the sole reference

spatial dataset, benchmarking the ABA data is essential to assess the robustness of the observed gene expression patterns across distinct experiments and technological platforms. In this manuscript we use “benchmarking” to refer to the assessment of replicability across independent datasets representing different experimental techniques. Obtaining replicable results across gene expression assays is notoriously challenging, so cross-platform, cross-dataset transcriptomics benchmarking has proved crucial since early transcriptome assays in the form of microarrays (Canales et al., 2006; Shi et al., 2006).

To address this need for spatial transcriptomics and cross-modality robustness in the brain, here we undertook a whole-brain benchmarking of the ABA via linking gene expression and anatomy. We analyzed a spatial gene expression dataset from one adult mouse brain collected using spatial transcriptomics (ST) (Ortiz et al., 2020) (see Section 3.3) alongside the ABA. ST is a spatially barcoded mRNA capture technique followed by sequencing read-out, while the ABA dataset is a collection of single-molecule in situ hybridization experiments across the whole-transcriptome (Ståhl et al., 2016; Lein et al., 2007). While benchmarking of the two datasets could be done on many scales, we chose to look across brains and across techniques with reference to named brain areas. This approach contains noise associated with the relative biases of each technique (different assays); experimental noise from tissue processing and alignment; biological variability (different brains); and variability from brain area segmentation and naming itself. Despite all these potential sources of noise, our approach combining spatial gene expression with brain area identity allows us to focus on biological conclusions that could be drawn from replicable spatial data. Not readily available with more technical approaches to benchmarking, our approach allowed us to pursue a biological question. We principally ask if canonical, anatomically-defined brain areas from the Allen Reference Atlas (ARA) can be assigned using gene expression alone and, in corollary, how well these assignments replicate across the ABA and ST datasets. We use an interpretable supervised learning framework for classification, where the target values are the ARA brain area labels and the features are the gene expression profiles for samples from across the whole brain (Figure 3.1a, b). We choose to use linear modeling to maintain easily interpretable models that can

be related to underlying biology.

Using this approach, we show that ARA labels are classifiable using gene expression, but that performance is higher in the ABA than ST. We further demonstrate that models trained in one dataset and tested in the opposite dataset do not reproduce classification performance bi-directionally. We then identify potential biological explanations for the difference in cross-dataset performance in classifying brain areas. Finally, we found that although an identifying gene expression profile can always be found for a given brain area, it does not generalize to the opposite dataset. In summary, within each dataset, canonical brain area labels were classifiable and meaningful in gene expression space, but replicability across these two very different assays of gene expression was not robust.

3.2 Results and discussion

3.2.1 Allen Reference Atlas brain areas are classifiable using gene expression alone

With the advent of new high-throughput capture technologies for spatial transcriptomics, we present, as is necessary for all new biological assays, a cross-technology assessment of generalizability in a well-characterized model system: the adult mouse brain. These new technologies allow, for the first time, the cross-platform assessment of canonical, atlas brain area subdivisions relative to gene expression at a whole-brain scale. Traditionally, parcellation of the mouse brain has depended on anatomical landmarks and cytoarchitecture, at times, including inter-region connectivity and molecular properties (Lein et al., 2007; Crick and Jones, 1993; MacKenzie-Graham et al., 2004). By enabling the relatively rapid and high-throughput collection of spatially-resolved, whole-transcriptome data in the adult mouse brain, these new spatial assays pave the way for a multi-modality assessment of canonical brain area labels. Specifically, in the present work we ask if brain areas from the Allen Reference Atlas (Lein et al., 2007) are classifiable using two spatial gene expression datasets: the Allen Institute's own in situ hybridization data (Lein et al., 2007; Ng et al., 2007) and a second dataset collected using Spatial Transcriptomics (Ståhl

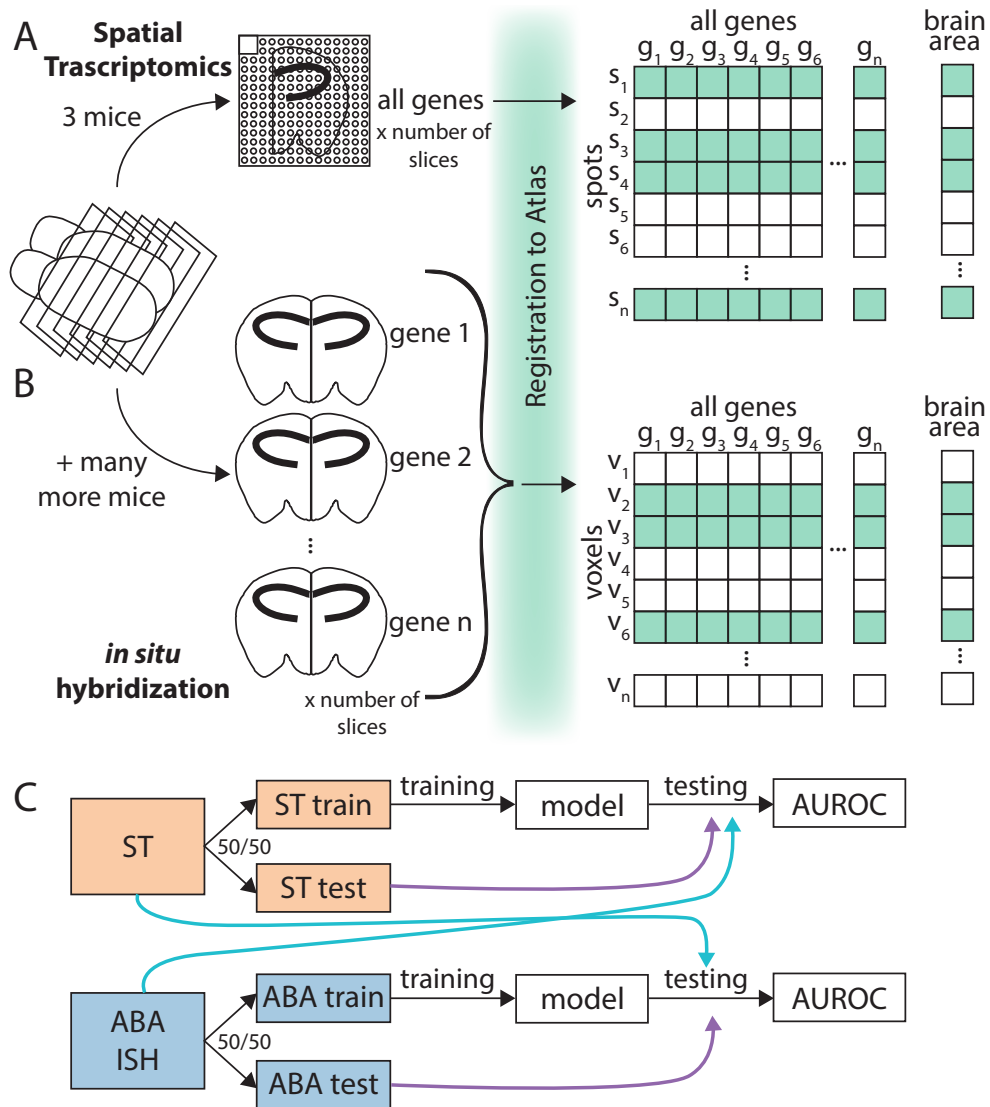


FIGURE 3.1: Collection and processing of spatial gene expression datasets. (A) Schematic depicting workflow of collecting whole brain spatial gene expression using Spatial Transcriptomics (ST). Illustration depicts sectioning of mouse brain, tissue from one hemisphere on one Spatial Transcriptomics slide, registration to Allen Reference Atlas, and a layout of the collected data. (B) Schematic depicting workflow of collecting Allen Institute’s whole brain spatial gene expression using in situ hybridization (ABA). Illustration depicts similar workflow to (A), but instead of Spatial Transcriptomics capturing all genes in one (three for this dataset) brain, there were many more mice used to collect the whole transcriptome dataset since each brain tissue slice can only be used to probe one gene. (C) Schematic illustrating classification schema. The ST dataset from (A) (orange) and ABA dataset from (B) (blue) were split into 50/50 train/test folds. The training fold was used for model building and the test fold for evaluating the trained model within dataset (purple arrow). Later analysis also applied models trained using the train fold of one dataset to the opposite dataset for testing (light blue arrow).

et al., 2016; Ortiz et al., 2020) (Figure 3.1a, b). After filtering, the ABA consists of 62,527 voxels (rows) with expression from 19,934 unique genes (columns) mapping to 569 non-overlapping brain area labels and the ST consists of 30,780 spots (rows) with 16,557 genes (columns) mapping to 461 brain area labels (see Section 3.3 for details). The ABA dataset consists of a minimum of roughly 3,260 brains, while the ST dataset is collected from 3 mice (Lein et al., 2007; Ortiz et al., 2020) (see Section 3.3). Comparing accuracy in classification of ARA brain areas across two technological platforms and datasets allows us to draw conclusions about spatial expression that are more likely to be biological and generalizable than subject to the technical biases of any one dataset.

To determine if we could more generally determine canonical brain areas from spatial gene expression, we first asked if we could do so within each of the two datasets independently. Given the known high correlation structure of gene expression (Eisen et al., 1998), we hypothesized that we could determine the brain area of origin of a gene expression sample using only a subset of the total genes. Fitting these criteria, we chose least absolute shrinkage and selection operator, or LASSO regression (Tibshirani, 1996). LASSO is a regularized linear regression model which minimizes the L1 norm of the coefficients (i.e. the sum of the absolute values of the coefficients). LASSO typically drives most coefficients toward zero, and thus leaves few genes contributing to the final model; LASSO in effect picks “marker genes” of spatial expression in the brain. We use LASSO in a supervised learning framework with a random 50/50 train-test split for two-class classification of all pairwise brain areas successively (Figure 3.1c) (see Section 3.3). The brain areas included here are non-overlapping and are the smallest brain areas present in the ARA naming hierarchy. We subsequently refer to these areas as leaf brain areas since they form the leaves of the tree-based representation of the ARA named brain areas (Lein et al., 2007). The performance of the test set classification is reported using the area under the receiver operating curve (AUROC). The AUROC can be thought of as the probability of correctly predicting a given brain region from its gene expression in a comparison with an out group (here, a different brain region) and is calculated by taking the predictions from the trained LASSO model and evaluating their correspondence with the known labels in

the test fold (see Section 3.3). For example, if ranking the samples by the LASSO predictions separates the samples from the two classes perfectly without being interspersed, we would get perfect classification with an AUROC of 1, while a score of 0.5 is random. More generally, in this manuscript, we say a brain area pair is classifiable with respect to each other to indicate a high performance in classification with an AUROC greater than 0.5 and generally closer to 1.

After preliminary filtering (see Section 3.3), we use this approach in both the ST and ABA to classify all the leaf brain areas against each of the others (461 ST areas; 560 ABA areas) (Figure 3.1c; see Section 3.3). ABA leaf brain areas are classifiable using LASSO ($\lambda=0.1$) from all other leaf brain areas using only gene expression data from (1) the ABA (mean AUROC = 0.996) (Figure 3.2a, Figure 3.3a) and from (2) the ST (mean AUROC = 0.883) (Figure 3.2b, 3.3b). These results are consistent across an additional, independent train/test fold split for both datasets (ABA mean AUROC = 0.996, correlation to first split, $\rho = 0.732$; ST mean AUROC = 0.882, correlation to first split, $\rho = 0.860$) (Figure 3.4a-d). As expected, performance falls to chance when brain area labels are permuted as a control (ABA mean AUROC = 0.510; ST mean AUROC = 0.501) (Figure 3.5a-d). Together, these results indicate that there is a set of genes whose expression level can be used to identify it and suggests that canonical brain area labels do reflect spatial patterning of gene expression assayed in both the ABA and ST datasets.

Since our task can be conceived as a multiclass classification problem, we asked if brain area classification performance could be improved using a true multiclass classifier. To test this question, we used the k-nearest neighbors (k-NN) algorithm which simply assigns the class identity of a test sample based on the majority class label (brain area) of its k closest neighbors in feature (here, expression) space. Using k-NN ($k = 5$), classification of leaf brain areas fell in ABA (mean AUROC = 0.695; Figure 3.6a) and ST (mean AUROC = 0.508; Figure 3.6b) (see Section 3.3). Given the lack of increase in performance and the preferability of our biologically interpretable approach, we choose to continue most analyses using LASSO.

We next asked if single gene marker selection strategies could outperform LASSO.

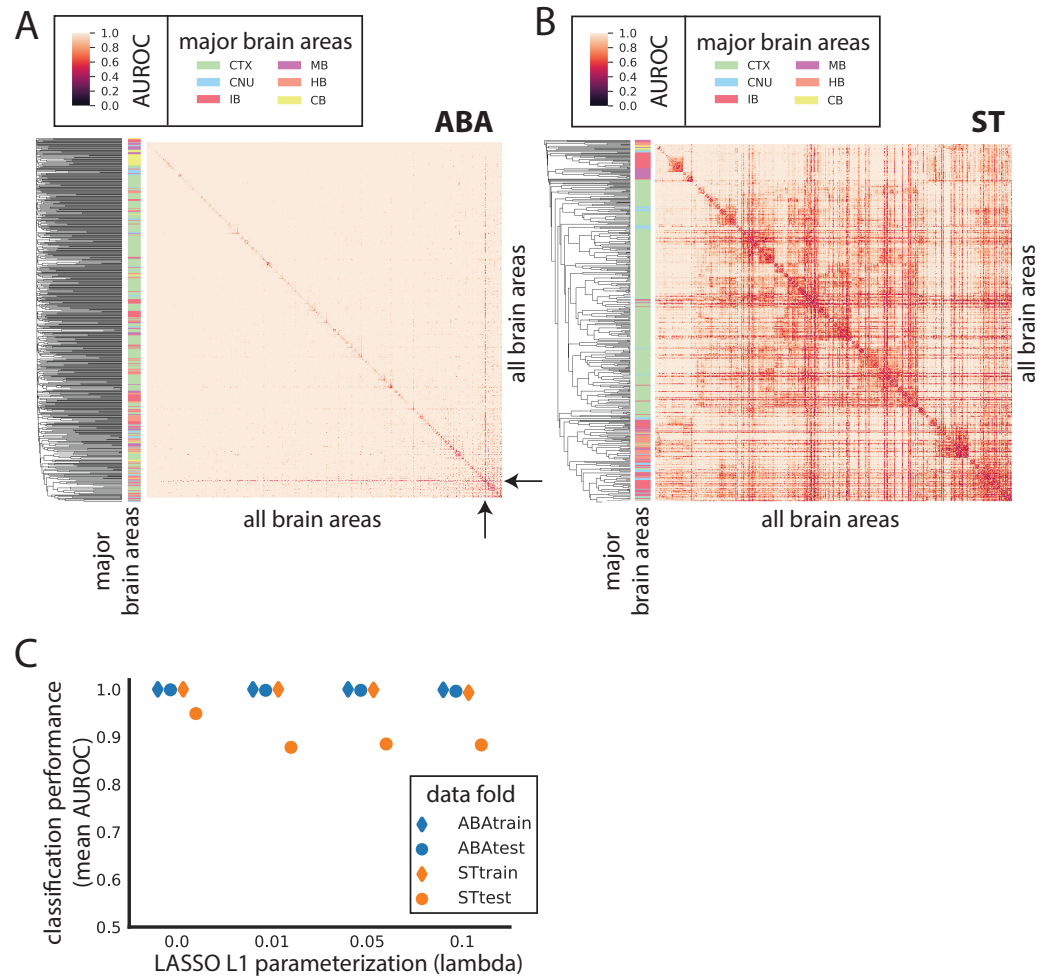


FIGURE 3.2: Canonical brain areas are classifiable using gene expression alone in the ABA and ST datasets. Heat map of AUROC for classifying leaf brain areas from all other leaf brain areas in (A) ABA and (B) ST using LASSO ($\lambda = 0.1$). Dendrograms on the far left side represent clustering of leaf brain areas based on the inverse of AUROC; areas with an AUROC near 0.5 get clustered together while areas with an AUROC near 1 are further apart. Color bar on the left represents the major brain structure that the leaf brain area is grouped under. These areas include: cortex (CTX), midbrain (MB), cerebellum (CB), striatum and pallidum (CNU), hindbrain (HB), and thalamus and hypothalamus (IB). (C) Average AUROC (y-axis) of classifying all brain areas from all other brain areas using LASSO across various values of λ (x-axis): 0, 0.01, 0.05, and 0.1 for ABA train (blue diamond), ABA test (blue dot), ST train (orange diamond), ST test (orange dot).

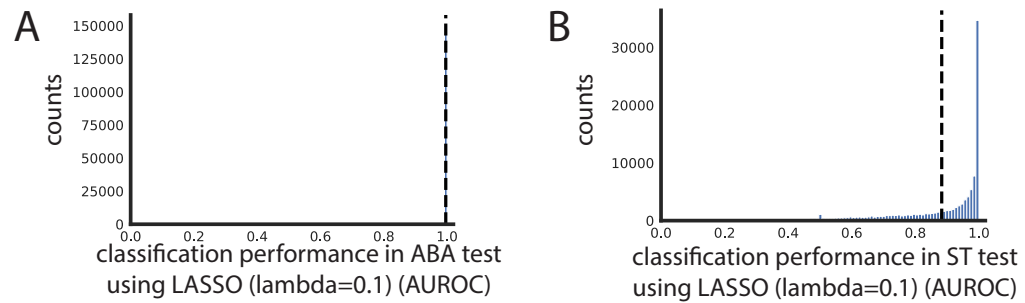


FIGURE 3.3: **Histogram visualization of LASSO ($\lambda = 0.1$) performance.** Histogram of classification performance of LASSO ($\lambda = 0.1$) in (A) ABA test fold and (B) ST. (A) and (B) represent the upper triangular of Figure 3.2a and Figure 3.2b respectively. Black dashed vertical line represents the mean.

Highlighting specific brain areas where such markers are known, we looked at classifying the CA2 of the hippocampus and arcuate hypothalamic nucleus with *Amigo2* and *Pomc*, respectively (Hitti and Siegelbaum, 2014; Toda et al., 2017; Laeremans et al., 2013). Following longstanding anatomical divisions of the mouse brain, the hippocampal sub-regions were re-defined in the mid 2000s using differences in gene expression (Lein, Zhao, and Gage, 2004; Lein et al., 2005). Follow-up to the early redefinitions found that while not exclusively expressed in the CA2, *Amigo 2* showed high expression levels in the CA2 (Laeremans et al., 2013). Indeed, in the CA2 of the hippocampus, *Amigo2* performs better than any other single gene in the ABA (*Amigo2* ABA AUROC = 0.920) and ST datasets (*Amigo2* ST AUROC = 0.612) (Figure 3.7a). However, classification of the CA2 using *Amigo 2* is still outperformed by the average performance of genes selected by LASSO. One of the major neuronal populations of the arcuate hypothalamic nucleus are the POMC-expressing neurons, shown to have a role in food intake and metabolism (Toda et al., 2017). In the arcuate hypothalamic nucleus, *Pomc* performance in the ABA (*Pomc* ABA AUROC = 0.993) and ST (*Pomc* ST AUROC = 0.910) is better than most other single genes and comparable or less than the average LASSO performance for each dataset Figure 3.7b). Given the comparable performance and, more importantly, since there are not such known markers for most brain areas, we again turned our attention to using LASSO for classifying brain areas.

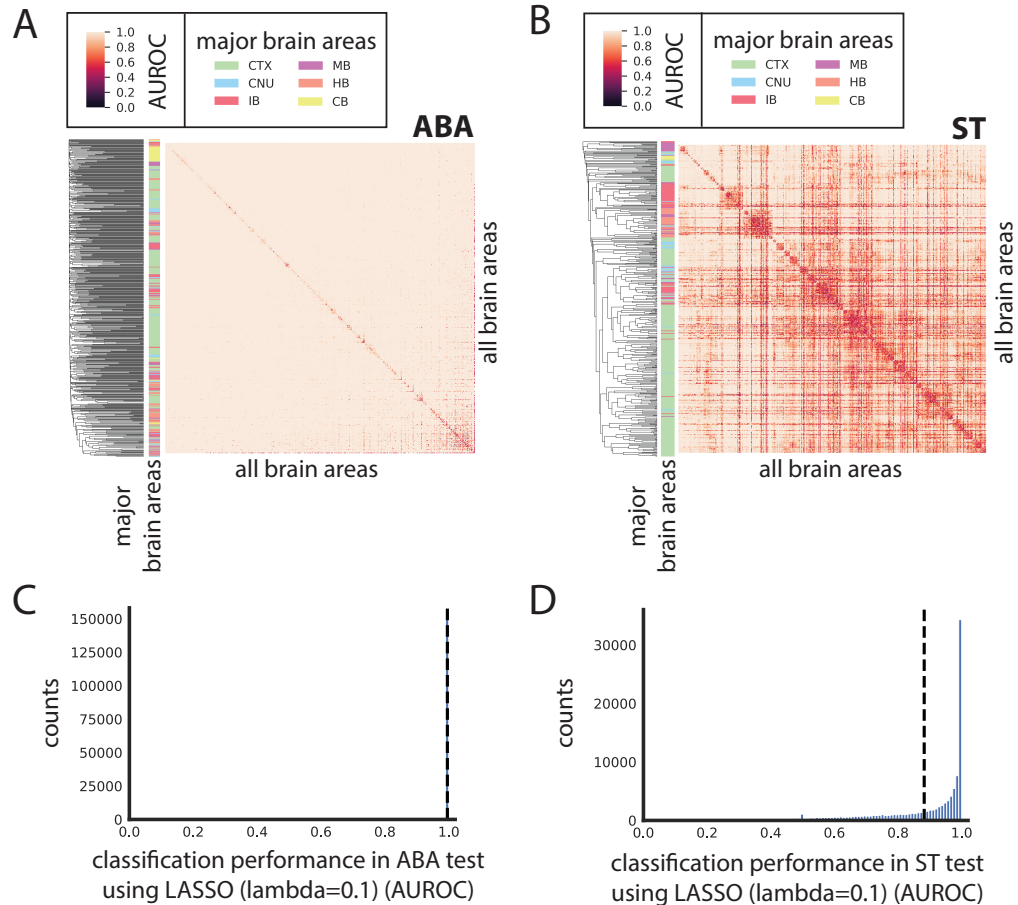


FIGURE 3.4: Additional visualization and verification of within dataset LASSO results with a new random train/test split. Heat map of AUROC for classifying leaf brain areas from all other leaf brain areas in (A) ABA and (B) ST using LASSO (lambda = 0.1) using a different random train/test split with a seed = 9 relative to Figure 3.2a-b. Dendrograms on the far left side represent clustering of leaf brain areas based on the inverse of AUROC; areas with an AUROC near 0.5 get clustered together while areas with an AUROC near 1 are further apart. Color bar on the left represents the major brain structure that the leaf brain area is grouped under. These areas include: cortex (CTX), midbrain (MB), cerebellum (CB), striatum and pallidum (CNU), hindbrain (HB), and thalamus and hypothalamus (IB). Histogram of classification performance of LASSO (lambda = 0.1) in (C) ABA test fold and (D) ST. (C) and (D) represent the upper triangular of (A) and (B) respectively. Black dashed vertical line represents the mean.

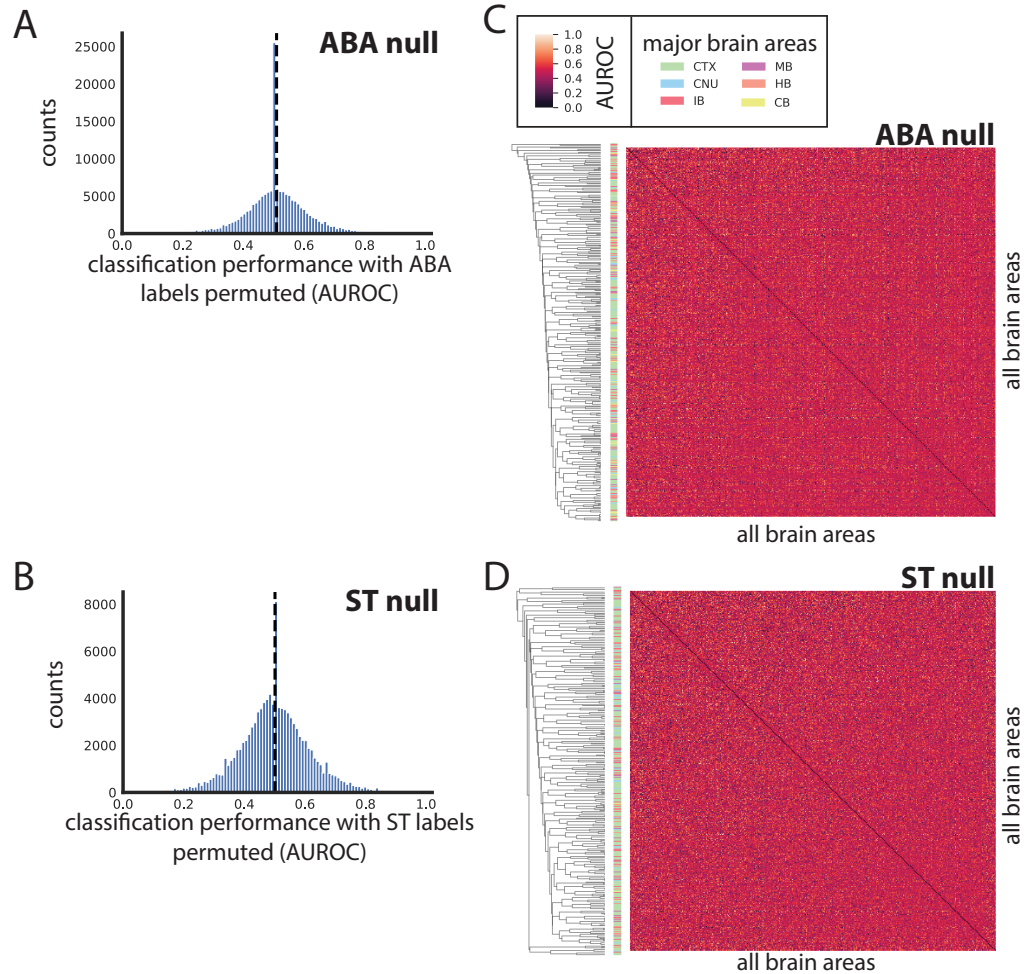


FIGURE 3.5: LASSO performance with permuted labels falls to chance. Upper triangular (A) histogram and (C) heatmap map of AUROC for classifying leaf brain areas from all other leaf brain areas in ABA using LASSO ($\lambda = 0.1$) when brain area labels are randomly permuted. (B) and (D) same as (A) and (C) respectively, but for ST with labels permuted. For heatmaps (C-D), dendrograms on the far left side represent clustering of leaf brain areas based on the inverse of AUROC; areas with an AUROC near 0.5 get clustered together while areas with an AUROC near 1 are further apart. Color bar on the left represents the major brain structure that the leaf brain area is grouped under. These areas include: cortex (CTX), midbrain (MB), cerebellum (CB), striatum and pallidum (GNU), hindbrain (HB), and thalamus and hypothalamus (IB).

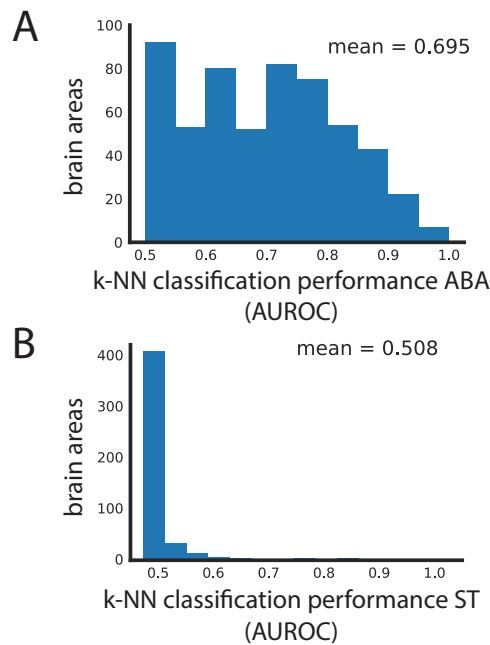


FIGURE 3.6: **Classifying brain areas using k-NN.** Distribution of performance (AUROC) of classifying each brain area using k-NN, a multiclass classifier, with default parameters ($k=5$) for (A) ABA and (B) ST. Mean AUROC is shown in the upper left of each plot.

Notably, performance using LASSO in the ABA is nearly perfect. That the classification in the ABA performs so well is striking, especially considering the potential loss of ISH-level resolution in the voxel representation of the ABA. For the median-performing pair of brain areas in ABA (median AUROC = 1), there is a threshold in classification that can be drawn where all instances of one class can be correctly predicted without any false positives (precision = 1). In contrast, in the ST, no such threshold can be found for the median-performing (median AUROC = 0.959) brain areas (average precision = 0.846) (see Section 3.3). Further, performance in the ABA is consistently higher than the ST across various parameterizations of LASSO (Figure 3.2c) (see Section 3.3). Despite the comparatively lower performance in the ST, clustering brain areas by AUROC shows brain areas belonging to the same major anatomical region grouping together (Figure 3.2b) (see Section 3.3). For example, most brain areas belonging to the cortex group together in the middle of the heat map (green bar on left) with a few interspersed areas. This grouping suggests that patterns of expression track with broad anatomical labels. Examining the relative

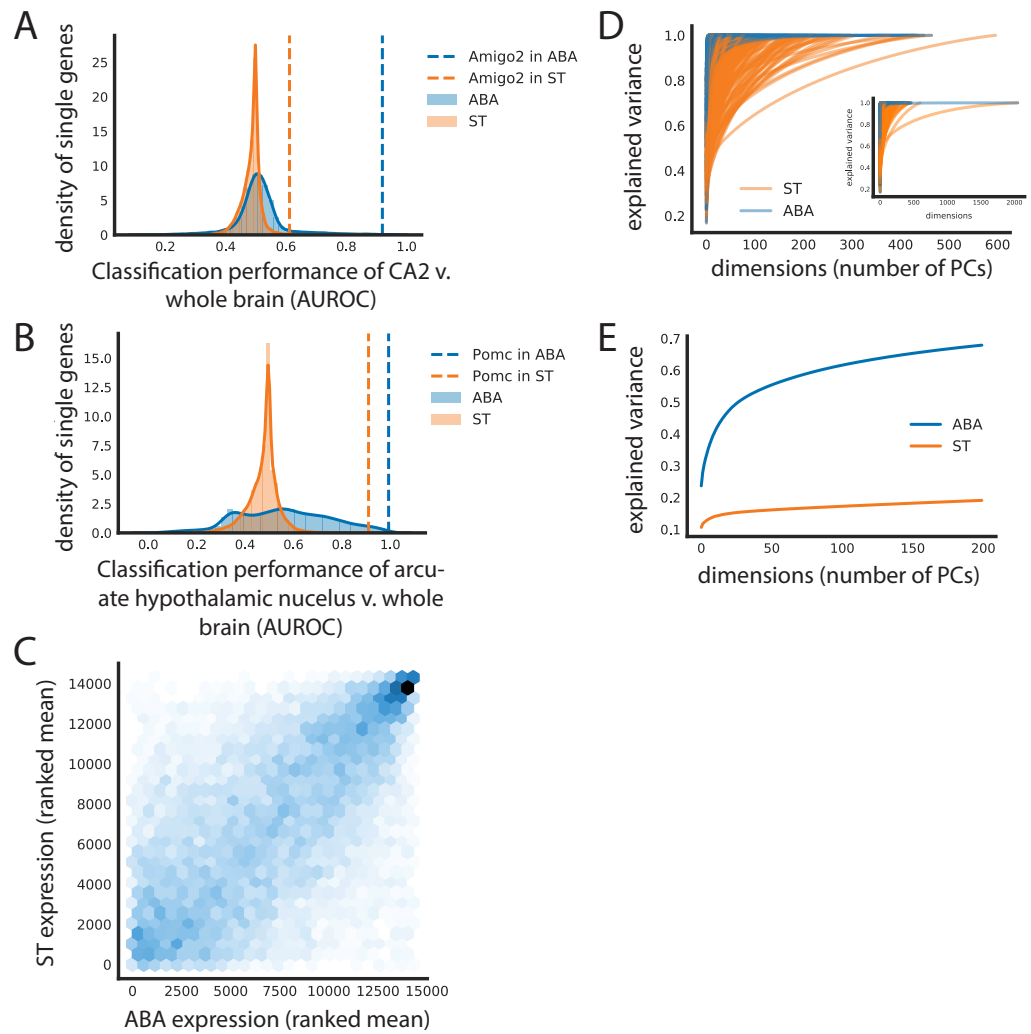


FIGURE 3.7: Classification using single genes, relative expression across datasets, and PCA. Distribution of classifying (A) CA2 and (B) arcuate hypothalamic nucleus against the rest of the brain using single genes. Distributions of all single genes shown for classification in ABA (blue) and ST (orange). Dashed lines represent the marker gene (A) Amigo2 or (B) Pomc for classification in ABA (blue) or ST (orange). (C) Relative expression between the ST and ABA datasets as a density plot. Expression is plotted as the ranked mean for each gene across all samples. (D) Cumulative explained variance curves for PCA in ST (orange) and ABA (blue). Each curve represents one leaf brain area. The total number of principal components per brain area is equal to the number of samples in that area. ABA areas that had more samples than ST are randomly down-sampled accordingly. For both datasets, the Caudoputamen, the largest region, is removed to allow visualization; full figure shown in inset plot. (E) Cumulative explained variance curves for 200 PC's in the whole ST (orange) and whole ABA (blue) datasets.

expression of genes that are assayed in both datasets, we see that ranked mean expression is comparable across the two datasets (Spearman's $\rho = 0.599$) (Figure 3.7c) suggesting that the observed difference in performance is not due to poorly detected genes being well-detected in the opposite dataset or vice versa.

Observing the nearly perfect performance in the ABA, we next hypothesized that this dataset may be more low-dimensional than suggested by its feature size and may contain many highly correlated features when compared to the ST dataset. We applied principal component analysis (PCA) in each brain area separately by subsetting the data by brain areas then calculating PCA in each of these subsets independently. Using this approach, we find that on average in individual brain areas, 2 PCs are enough to summarize 80% of the variance per brain area in ABA versus 21 PCs in ST (Figure 3.8a; Figure 3.7d) (see Section 3.3). In other words, within each brain area in the ABA, many genes are highly co-expressed. Zooming out to the whole brain, using 200 PCs captures nearly 70% of the variance in ABA compared to nearly 20% in ST (Figure 3.7e). Further, gene-gene co-expression across the whole dataset is on average higher in the ABA (gene-gene mean Spearman's $\rho = 0.525$) than the ST (gene-gene mean Spearman's $\rho = 0.049$) (Figure 3.8e). The perfect performance, low-dimensionality on a per brain area basis, and high co-expression all support the idea that although there is meaningful variation in the ABA, it can be captured in few dimensions. In summary, canonical ABA brain areas are classifiable from each other using gene expression alone, but performance is likely inflated in the ABA.

An aside of note is that in the ABA the one brain area that is consistently lower performing when classified against most other brain areas is the Caudoputamen (mean AUROC = 0.784) (Figure 3.2a, black arrows). In the ST, the Caudoputamen is not the lowest performing area, but also has a low mean AUROC (AUROC = 0.619) relative to the other brain areas in ST. In both datasets, the Caudoputamen is the largest leaf brain area composed of the most samples (ABA CP number of voxels = 3012 vs. an average of 85.6 voxels; ST number of spots = 2051 vs. an average of 57 spots). The Caudoputamen is similarly large in other rodent brain atlases, reflecting its lack of cytoarchitectural features

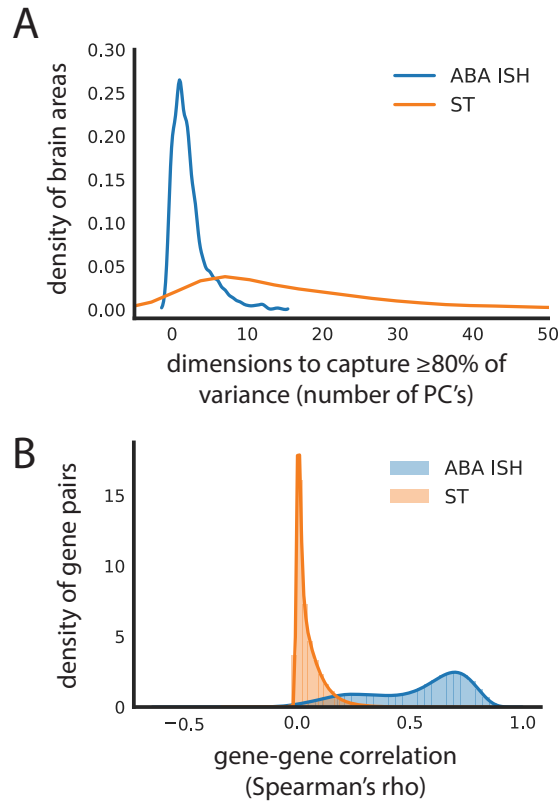


FIGURE 3.8: **Internal data structure of ABA and ST datasets.** (A) Number of principal components to capture at least 80% of variance of genes in each of the leaf brain areas after applying PCA to ABA (blue) and ST (orange). ABA brain areas that are larger than ST are randomly down-sampled to have the same number of samples as ST prior to applying PCA. (B) Gene-gene correlations calculated as Spearman's rho between all pairwise genes across the whole dataset for both the ABA (blue) and ST (orange) independently.

(Hintiryan et al., 2016). We hypothesized that its relatively larger size could mean that it consists of transcriptomically disparate sub-sections that are not captured with canonical ARA labelling. Though not an outlier, we do observe that the mean sample correlation for the Caudoputamen in both the ST (mean Pearson's $r = 0.727$) and ABA (mean Pearson's $r = 0.665$) is slightly lower than the mean in either case (ST mean Pearson's $r = 0.783$; ABA mean Pearson's $r = 0.696$) (Figure 3.9a). More generally, however, we observe that there is no relationship between size and performance across brain regions (Figure 3.9b,c). In addition to being an outlier in terms of size, the Caudoputamen is the dorsal part of the striatum which encompasses many different functional subdivisions evident through the various cortico-striatal projections (Hintiryan et al., 2016). Together with the low classification performance of the Caudoputamen using gene expression, this reflects the shortcomings of the ARA Caudoputamen label and the likely need to sub-divide the Caudoputamen functionally.

3.2.2 Cross-dataset learning of Allen Reference Atlas brain areas

Cross-dataset performance is not bi-directional Given the low-dimensionality and the near perfect brain area classification performance in the ABA relative to the ST dataset, we hypothesized that the performance of the LASSO models was artificially inflated in the ABA. To explore this hypothesis, we characterized whether LASSO models trained in one dataset would generalize to the opposite dataset (Figure 3.1c, light blue arrows). For this step we further filtered for (1) 445 leaf brain areas that were represented with a minimum of 5 samples in each dataset and for (2) 14,299 overlapping genes (see Section 3.3). In this section, we filtered within-dataset analyses to match this set of genes and leaf areas to maintain a parallel evaluation. LASSO-regularized linear models ($\lambda = 0.1$) trained on ST had a similar within-dataset performance (held out test fold, mean AUROC = 0.884) and cross-dataset performance (ABA, mean AUROC = 0.829) (Figure 3.10a,b), but the reverse is not true. The performance in classifying pairwise leaf brain areas using LASSO models trained in the ABA (held out test fold, mean AUROC = 0.997) falls when testing in the ST (mean AUROC = 0.725) (Figure 3.10a,c). These results are consistent across an additional random train/test split for both (1) ST (within-dataset ST

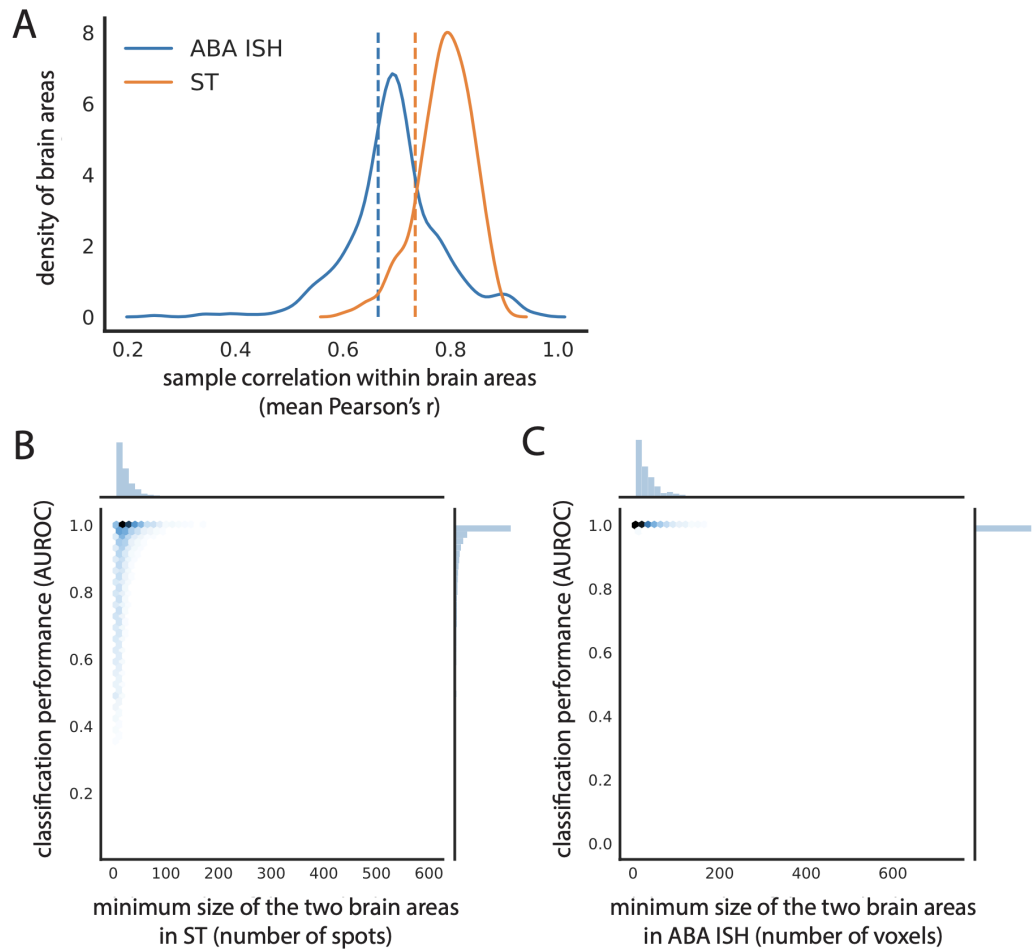


FIGURE 3.9: Sample correlation within brain areas and relationship between size and classification performance. (A) Distribution of sample correlation within each of the leaf brain areas for ABA (blue) and ST (orange). Vertical dashed line represents the mean for the correspondingly colored distribution. Test set AUROC in (B) ST and (C) ABA as a function of the number of samples per brain area. The minimum of the two brain areas involved in classification is shown.

test mean AUROC = 0.884; correlation to first split, $\rho = 0.735$) to ABA (ST to ABA cross-dataset mean AUROC = 0.831; correlation to first split, $\rho = 0.718$) (Figure 3.11a,b) and (2) for ABA (within-dataset ABA test mean AUROC = 0.997; correlation to first split, $\rho = 0.780$) to ST (ABA to ST cross-dataset mean AUROC = 0.722; correlation to first split, $\rho = 0.816$) (Figure 3.11a,c). These results show that the ST dataset is more generalizable to the opposite dataset than the ABA. Additionally, this discrepancy in cross-dataset performance suggests that the high performance within the ABA is driven by a property of that dataset not present in the ST (see discussion in Section 3.4).

Given this difference in cross-dataset performance, we next explored if correcting for batch effects improves cross-dataset classification performance. We treated each of the two datasets as a batch. Batches within each dataset are not clear, particularly in the ABA where batches might arise independently for each gene, which are sampled as an individual experiment by design of single molecule in situ hybridization. After batch correction between the datasets (see Section 3.3), there is virtually no difference in the mean AUROC for either cross-dataset comparison (ABA held-out test fold mean AUROC = 0.997; ABA to ST mean AUROC = 0.725; ST held-out test fold mean AUROC = 0.884; ST to ABA mean AUROC = 0.829). Looking at individual brain area pairs, there are some minor differences between un-corrected and corrected classification performance with the largest being for the ST within-dataset held out test fold (mean absolute difference between corrected and un-corrected = 0.001). Hypothesizing that the much larger ABA dataset could be driving the batch correction and thus showing very little difference between corrected and un-corrected performance, we down-sampled the ABA to have the same sample size as the ST. Filtering, as before, after down-sampling left us with 414 brain areas. Compared to un-corrected performance when filtering for the same brain areas there are very small differences in the mean AUROCs (see S1 Table in Appendix C). Visualizing the two datasets in principal component space suggests that batch correction may not have much effect since there are no obvious global differences relative to one another (Figure 3.12a-f).

We next ask if the high performance seen within ABA that is lost when models built in the ABA are evaluated in the ST is specific to the LASSO method or a more general

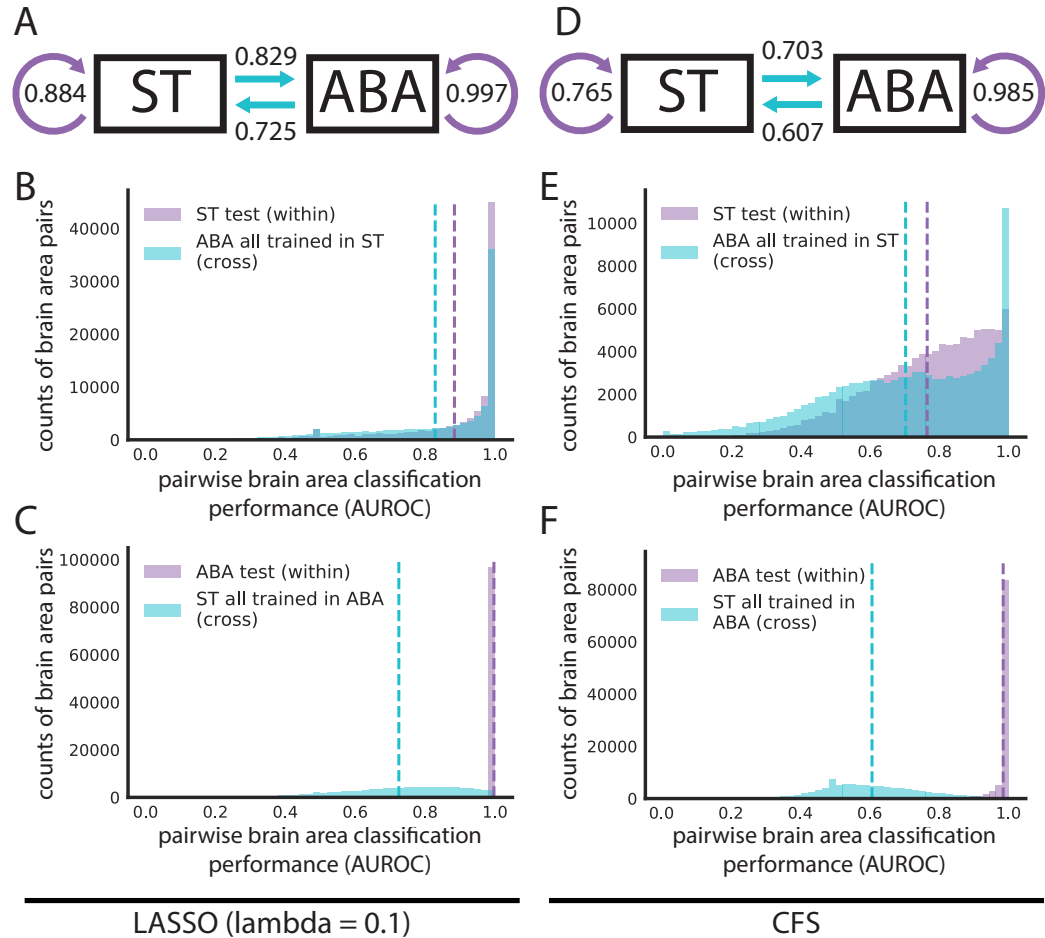


FIGURE 3.10: Cross-dataset learning shows that models do not generalize bi-directionally. (A and D) Models trained with overlapping genes and brain areas between ST and ABA datasets are evaluated within dataset on the test fold and across dataset on the entire opposite dataset as illustrated in Figure 3.1C. Summary diagrams showing mean AUROC for within-dataset test set performance (purple arrow) and cross-dataset performance with models trained in the opposite dataset (light blue arrow) for (A) LASSO (lambda = 0.1) and (D) CFS. Distributions of AUROCs for within- (purple) and cross-dataset (light blue) performance for (B) LASSO (lambda = 0.1) trained in ST, (C) LASSO (lambda = 0.1) trained in ABA, (E) CFS trained in ST, and (F) CFS trained in ABA. In all four plots, dashed vertical lines represent the mean of the corresponding colored distribution.

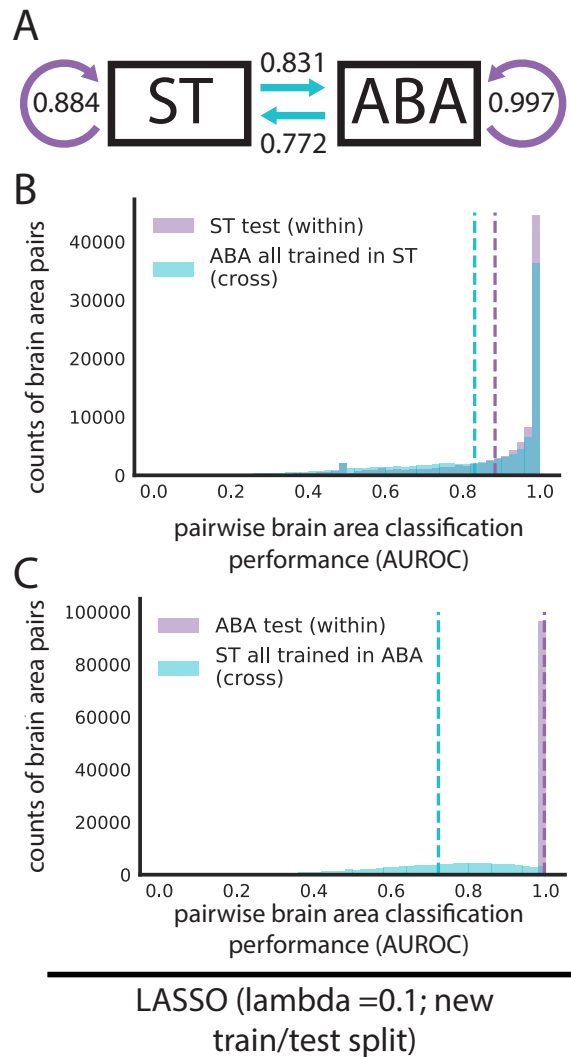


FIGURE 3.11: Additional verification of cross dataset LASSO results with a new random train/test split. (A) Models trained with overlapping genes and brain areas between ST and ABA datasets are evaluated within dataset on the test fold and across dataset on the entire opposite dataset as illustrated in Figure 3.1C. Summary diagrams showing mean AUROC for within dataset test set performance (purple arrow) and cross dataset performance with models trained in the opposite dataset (light blue arrow) for (A) LASSO (lambda = 0.1) with a new random train/test split (seed = 9) relative to Figure 3.10a-c. Distributions of AUROCs for within (purple) and cross dataset (light blue) performance for (B) LASSO (lambda = 0.1) trained in ST with new train/test split and (C) LASSO (lambda = 0.1) trained in ABA with new train/test split. In both plots, dashed vertical lines represent the mean of the corresponding colored distribution.

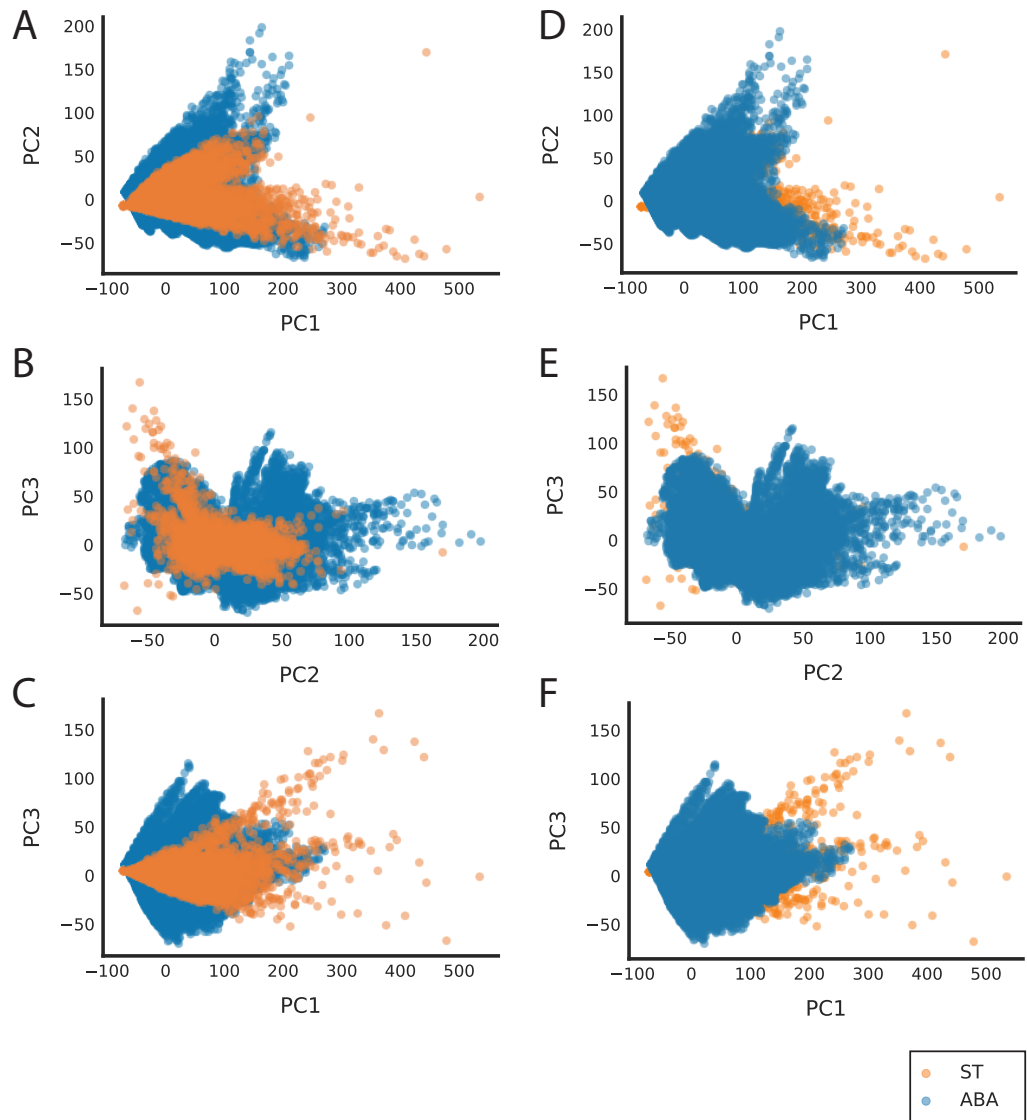


FIGURE 3.12: **Visualization of the ABA and ST datasets together in low-dimensional space.** Plots showing ABA (blue) and ST (orange) datasets visualized together in low-dimensional space after dimensionality reduction with PCA. (A-C) show ST plotted on top of ABA, while (D-F) show the same PCs with ABA plotted on top of ST. Plots show (A,D) PC1 v. PC2, (B,E) PC2 v. PC3, and (C,F) PC1 v. PC3.

feature of the data. To assess the data more directly, we used a second simpler method, correlation-based feature selection (CFS). CFS eliminates model building and simply picks features (genes) that are uncorrelated (Hall, 1999) (see Section 3.3). In this way, CFS parallels LASSO which implicitly picks uncorrelated feature sets when minimizing its L1 regularized cost function that penalizes additional features.

Using CFS, we picked 100 randomly seeded feature sets for pairwise comparisons of leaf brain areas (see Section 3.3, Figure 3.13a,b). We then took the single best performing feature set from the train set and evaluated its performance on both the held-out test set and cross-dataset. We did this in both directions, training on both ST and ABA as with LASSO above. CFS can accurately classify pairwise leaf brain areas in both the ST (test set mean AUROC = 0.765) and in the ABA (test set mean AUROC = 0.985) (Figure 3.10d-f). As with LASSO, classification in ABA with CFS is on average better performing than in ST. Again, following a similar trend as LASSO, the difference in mean cross-dataset performance going from the ST test set to the ABA (difference in mean AUROC = 0.052; mean ST to ABA cross-dataset AUROC = 0.703) is smaller than the reverse (difference in mean AUROC = 0.378; mean ABA to ST AUROC = 0.607) (Figure 3.10d-f). Altering our analysis approach by averaging the 100 CFS feature sets, we again see a similar pattern in cross-dataset performance (ST to ABA difference in mean AUROC = 0.062; ABA to ST difference in mean AUROC = 0.381) (Figure 3.13c-e). These CFS results indicate that the observed high performance of classification within the ABA and lack of generalization to the ST is not driven by our choice of model. In summary, across both techniques, marker genes can be found to classify pairwise leaf brain areas from each other, but they often do not generalize to the opposite dataset.

The sagittal subset of the ABA is the most distinct

With only two datasets it is impossible to distinguish whether the above lack of bi-directionality in cross-dataset learning is driven by (1) the ST being more generalizable or (2) a lack of information in ST that is critical to the high classification performance within ABA. To begin to address this, we took advantage of the separability of the ABA dataset into two distinct datasets: coronal and sagittal. The Allen Institute collected duplicates of

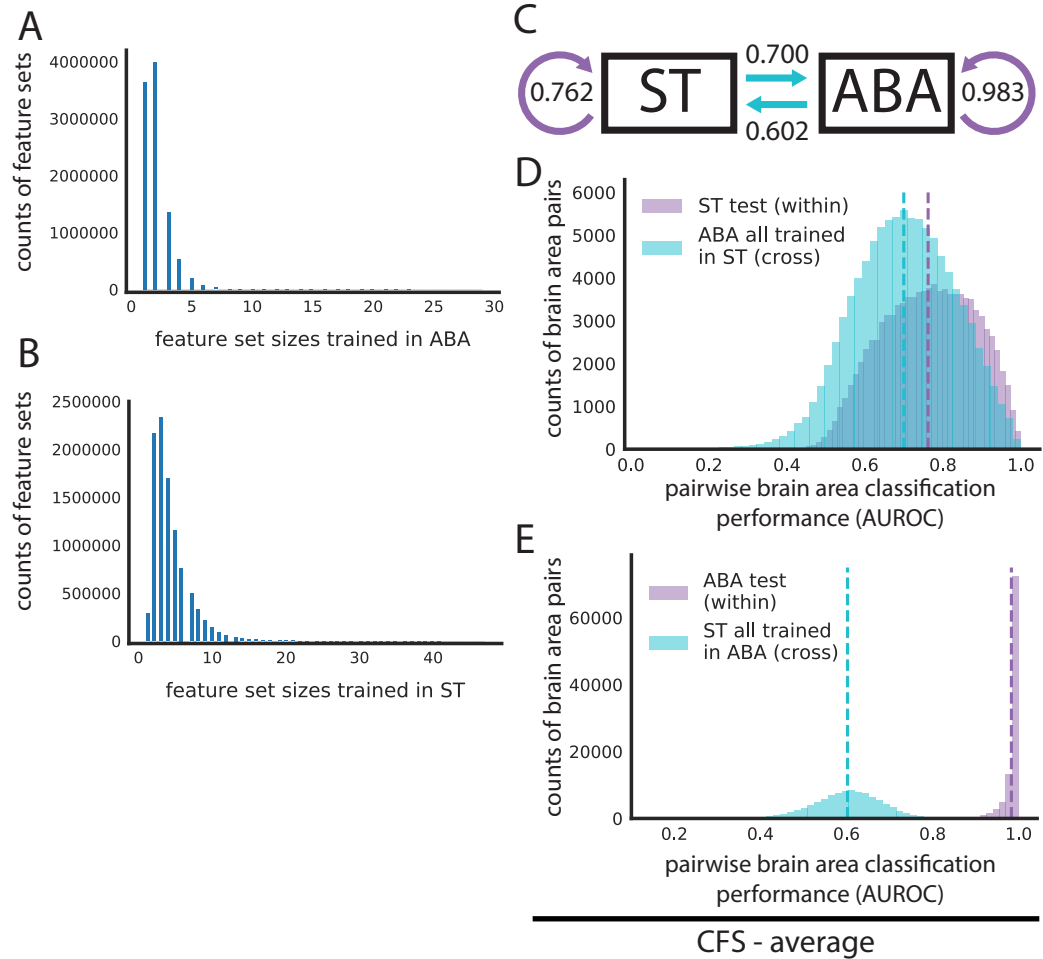


FIGURE 3.13: Feature set sizes for correlation-based feature selection (CFS) and cross-dataset results for CFS with averaging across feature sets. Distribution of correlation-based feature selection feature set sizes for (A) ABA and (B) ST. Models trained with overlapping genes and brain areas between ST and ABA datasets are evaluated within dataset on the test fold and across dataset on the entire opposite dataset as illustrated in Figure 3.1C. (C) Summary diagrams showing mean AUROC for within dataset test set performance (purple arrow) and cross dataset performance with models trained in the opposite dataset (light blue arrow) using the average of 100 feature sets chosen with CFS. Distributions of AUROCs for within (purple) and cross dataset (light blue) performance for 100 averaged CFS picked gene sets trained (D) in ST and (E) in ABA. In both plots, dashed vertical lines represent the mean of the correspondingly colored distribution.

many genes; roughly 4,000 genes were collected across both the coronal and sagittal planes of slicing. With these two datasets alongside the ST, we further filtered for 3,737 overlapping genes across the same 445 leaf brain areas (see Section 3.3) and computed all pairwise combinations of cross-dataset learning. Notably, using LASSO ($\lambda = 0.1$), training on ST outperforms either plane of ABA in cross-dataset predictions: (1) ST to ABA coronal (mean AUROC = 0.888) performs better than ABA sagittal to ABA coronal (mean AUROC = 0.775) and (2) ST to ABA sagittal (mean AUROC = 0.671) performs better than ABA coronal to ABA sagittal (mean AUROC = 0.663) (Figure 3.14a). Further, the performance of models trained in ABA coronal to ABA sagittal (mean AUROC = 0.663) and ST to ABA sagittal (mean AUROC = 0.671) is lower than that of ABA coronal and ST to each other (ST to ABA coronal mean AUROC = 0.888; ABA coronal to ST mean AUROC = 0.796) (Figure 3.14a). This shows that the ABA coronal and ST are able to generalize to each other better than to the ABA sagittal. Across parametrizations of our model, the sagittal subset of the ABA continues to be the most distinct of the three datasets with the least generalizability (Figure 3.15a,b). To evaluate whether our selection of λ had a significant impact on these findings, we looked at a subset of brain areas with larger sample sizes (minimum of 100) to allow dynamic LASSO hyperparameter fitting and compared it with a fixed hyperparameter ($\lambda = 0.1$) in the same brain areas. This showed that performance was very similar between the two (Figure 3.15c,d) (see Section 3.3).

The relative distinctness of the ABA sagittal dataset could be driven by its sparsity—consisting of zeros for more than half of the dataset (53.9%) compared to only 7.5% zeros in the coronal subset. LASSO is able to find a robust set of marker genes within the ABA sagittal that does not reflect the best possible set of genes in the less sparse ABA coronal and ST. While the coronal subset of the ABA was curated for genes showing spatial patterning (Lein et al., 2007), the subset of the sagittal genes in this analysis contains only those also present in the coronal set. So, the lack of generalizability of the sagittal subset is particularly suggestive of technical experimental or downstream processing issues rather than the absence of spatial patterning in the genes themselves.

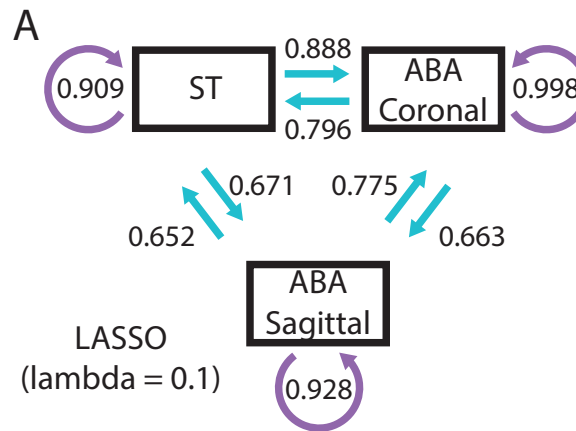


FIGURE 3.14: **Three-way cross-dataset LASSO shows that the sagittal subset of the ABA is the most distinct.** (A) Summary diagram showing mean AUROCs using LASSO ($\lambda = 0.1$) for separating out the two planes of slicing in the ABA and treating them alongside the ST dataset as three different datasets for cross-dataset learning. In all three summary diagrams (A, D, G), cross-dataset arrows originate from the dataset that the model is trained in and point to the dataset that those models are tested in.

3.2.3 Distance in semantic space, but not physical space provides a potential explanation for cross-dataset performance

Since the ARA brain areas are organized into a hierarchical tree-like structure based on biology (Lein et al., 2007), we hypothesized that the semantic distance of any two pairwise brain areas in this tree could provide an explanation for the cross-dataset performance of classifying samples from the same two areas. To investigate this, we used the path length of traversing this tree to get from one brain area to the second area as the measure of distance in the tree (see Section 3.3). For the performance of classifying brain areas in both the ST and ABA when trained in the opposite dataset (LASSO, $\lambda = 0.1$), we see an increase in performance (ST to ABA mean AUROC= 0.690 increases to mean AUROC = 0.912; ABA to ST mean AUROC = 0.655 increases to mean AUROC = 0.756) as the semantic distance increases from the minimum value of 2 to the maximum of 15 (Figure 3.16a,b). As expected, the corresponding increase in performance and semantic distance holds across parameterizations of our linear model (Figure 3.17a-d). A high AUROC here indicates that the two brain areas are transcriptionally distinct, while an AUROC near 0.5 indicates that they are similar. So, this result implies that distance in

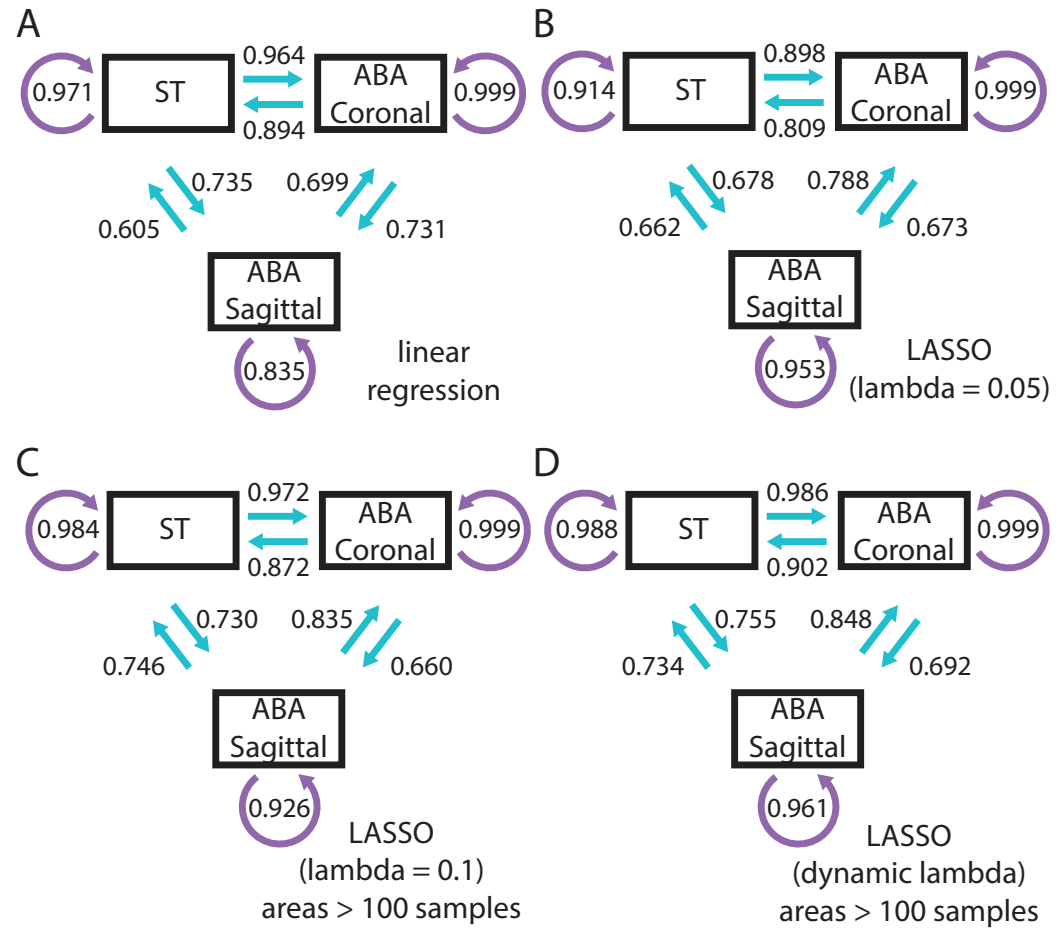


FIGURE 3.15: Summary plots for cross dataset analysis of ST, ABA coronal, and ABA sagittal with various parameterizations. Separating out the two planes of slicing in the ABA and treating them alongside the ST dataset as three different datasets for cross-dataset learning. Summary diagram showing mean AUROCs using (A) linear regression and (B) LASSO (lambda = 0.05). (C, D) Same cross dataset analysis as (A, B), but looking only at brain areas with at least 100 samples for (C) fixed lambda = 0.1 and (D) dynamically fitted lambda for each brain area pair. In all four summary diagrams, cross dataset arrows originate from the dataset that the model is trained in and point to the dataset that those models are tested in.

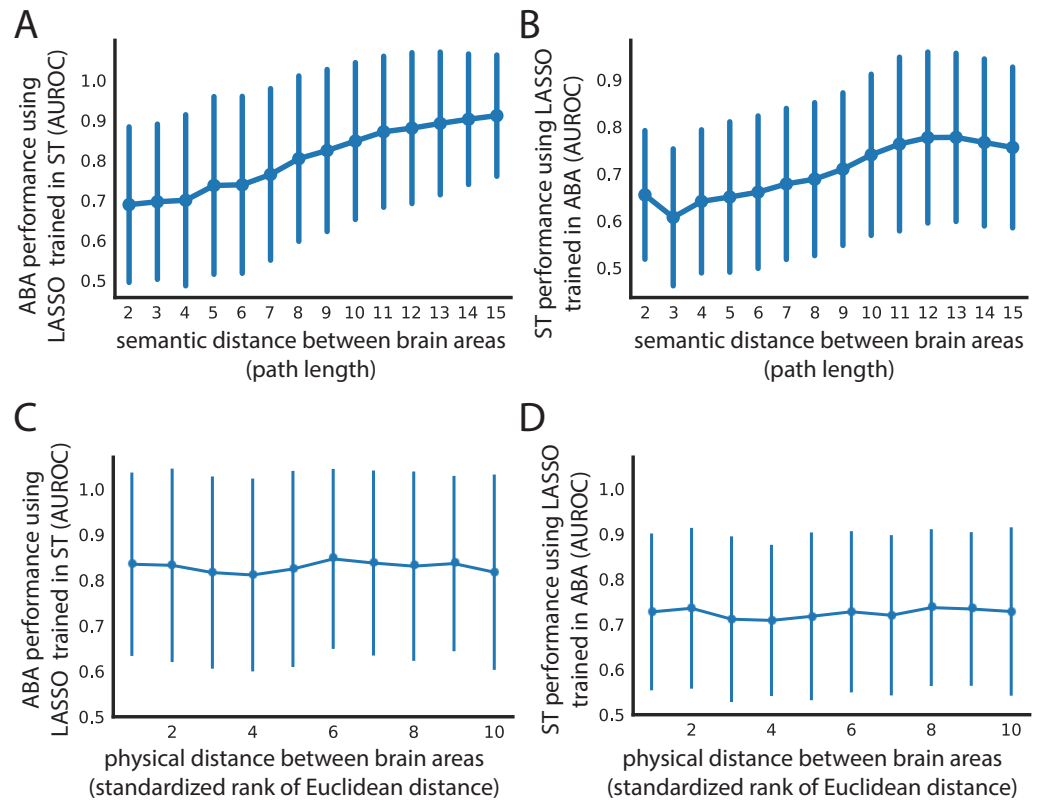


FIGURE 3.16: Spatial expression patterns reflect distance in semantic space, but not physical distance in the brain. Cross-dataset AUROCs (x-axis) of classifying all leaf brain areas from all other leaf brain areas for (A) ABA using LASSO ($\lambda = 0.1$) trained in ST and (B) ST using LASSO ($\lambda = 0.1$) trained in ABA as a function of path length (x-axis) in the ARA naming hierarchy between the two brain areas being classified. The same AUROCs (y-axis) from (A) and (B) shown in (C) and (D) respectively as a function of minimum Euclidean distance between the two brain areas in the ARA (x-axis). Euclidean distance on the x-axis is binned into deciles for visualization. All four plots show mean AUROCs (points) with standard deviation (vertical bars).

semantic space defined by the ARA reflects distance in expression space. This suggests that differences in classification performance are likelier to reflect real differences in gene expression between brain areas and not just large-scale gradients of expression present in the brain (Fornito, Arnatkevi, and Fulcher, 2018).

To further understand the relationship between performance and semantic distance, we next investigated pairs of brain areas with extreme AUROCs at the minimum and maximum semantic distances. We were especially interested in this given the distribution of AUROCs for each distance (Figure 3.16a, b). Similarly, at the smallest semantic distance

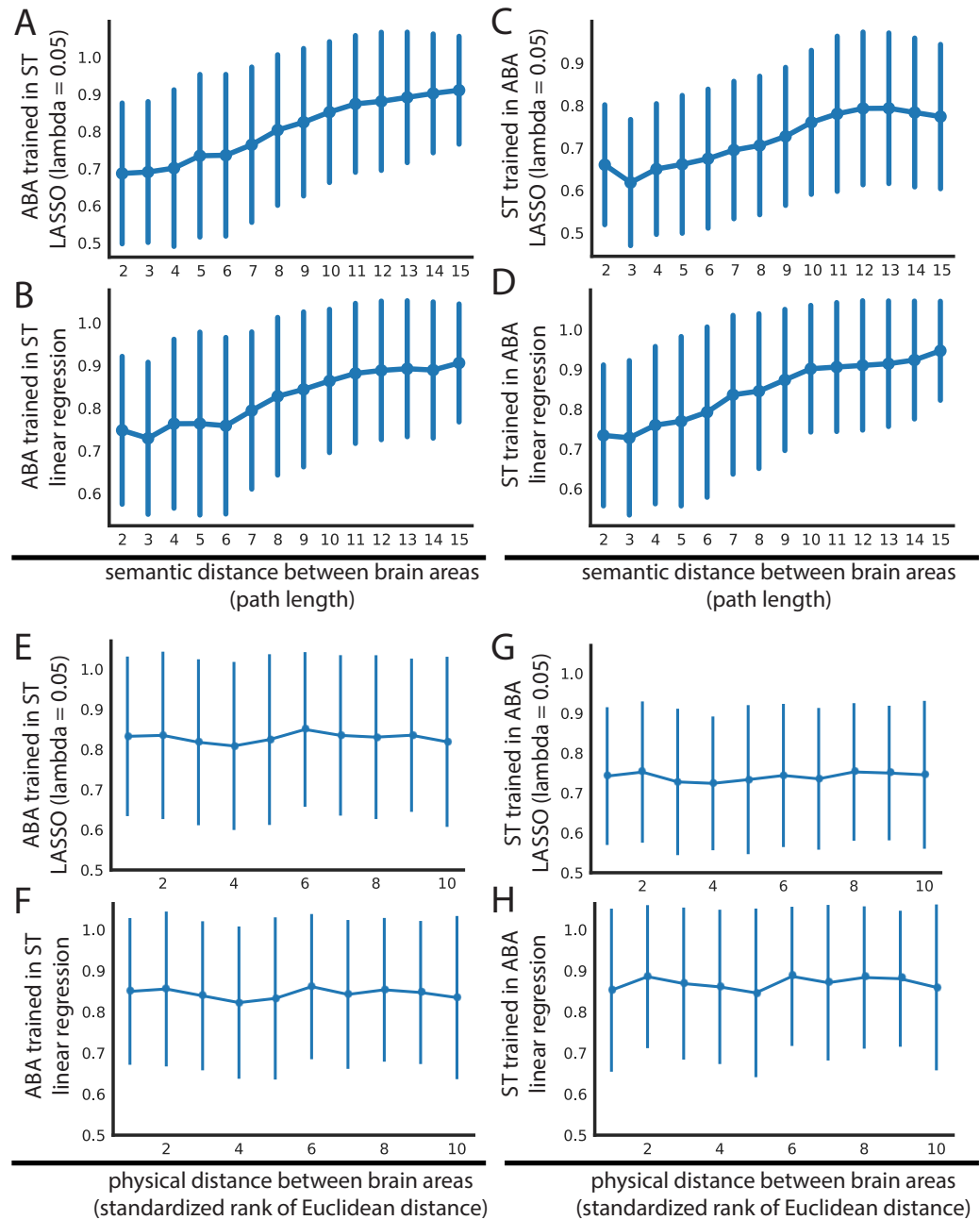


FIGURE 3.17: Comparison of path length and Euclidean distance to LASSO performance for various parameterizations of LASSO. Cross-dataset AUROCs (x-axis) of classifying all leaf brain areas from all other leaf brain areas for (A) ABA using LASSO ($\lambda = 0.05$) trained in ST, (B) ABA using linear regression trained in ST, (C) ST using LASSO ($\lambda = 0.05$) trained in ABA, and (D) ST using linear regression trained in ABA as a function of path length (x-axis) in the Allen Reference Atlas (ARA) naming hierarchy between the two brain areas being classified. The same AUROCs (y-axis) from (A-D) shown in (E-H) respectively as a function of minimum Euclidean distance between the two brain areas in the ARA (x-axis). Euclidean distance on the x-axis is binned into deciles for visualization. All plots show mean AUROCs (points) with standard deviation (vertical bars).

of 2, in both ABA and ST trained in the opposite dataset there is a spread in classification performance (see S2 Table, S3 Table in Appendix C). In both datasets, these brain area pairs involve different cortical layers of the same cortical area. The ARA hierarchy is organized such that within one cortical area, all the layers will have a semantic distance of 2 between each other. So, a pair of brain areas with a high AUROC and semantic distance of 2 often involves two non-neighboring layers of a cortical area (i.e. primary auditory cortex layer 6b and layer 4 in ST trained in ABA) (see S3 Table in Appendix C). This trend is in line with our expectation as cortical layers are known to have distinct expression profiles driven in part by distinct cell types (Yao et al., 2020; Yao et al., 2021; Codeluppi et al., 2018; Poulin et al., 2016). Alternatively, a pair of brain areas with an AUROC near 0.5 and a semantic distance of 2 can involve two neighboring layers of a cortical area (i.e. primary visual area layer 6a and layer 6b in ST trained in ABA) (see S3 Table in Appendix C). This too is not surprising because, despite distinctness in cortical layer expression, we expect some overlap between physically neighboring areas in terms of expression profiles due to errors introduced in sampling and in registration to the reference atlas. Together, these examples illustrate one way in which semantic distance is not synonymous to physical distance.

Since semantic distance does not perfectly capture the actual distance between brain areas, we next looked at classification performance as a function of physical distance directly. Specifically, we asked: Is performance in classifying pairwise leaf brain areas cross-dataset being driven by physical proximity/distance alone? Cross-dataset performance was examined with respect to the minimum Euclidean distance between the two brain areas in the ARA (see Section 3.3). There is no trend between physical distance and AUROCs from either cross-dataset assessment using LASSO models trained in the opposite dataset (ABA to ST Pearson's $r = -0.026$; ST to ABA Pearson's $r = 0.056$) with the mean performance remaining similar at the minimum (ABA to ST mean AUROC = 0.651; ST to ABA mean AUROC = 0.702) and maximum distance (ABA to ST mean AUROC = 0.697, change in AUROC = +0.046; ST to ABA mean AUROC = 0.690, change in AUROC = -0.012) (Figure 3.16c, d) (see Section 3.3). Across model parameterizations, there

is similarly no relationship between distance and performance (Figure 3.17e-h). This result alongside the positive relationship seen between performance and semantic distance, shows that spatial patterning of gene expression captures canonical brain area labels and is not merely composed of differences in large-scale gradients.

3.2.4 Finding a uniquely identifying gene expression profile for individual brain areas

Within one dataset a gene expression profile can uniquely identify one brain area, but it does not generalize to the opposite dataset

Thus far, we have focused on the classification of leaf brain areas from other leaf brain areas. However, this does not determine if we can uniquely identify a given brain area from the whole brain using gene expression. If possible, this could yield a set of marker genes to identify brain areas at their smallest parcellation for future neuroscience experiments. To tackle this, we trained linear models for one leaf brain area against the rest of the brain (one versus all) and tested that same model's performance in classifying the same leaf brain area against all others (one versus one across all leaf brain areas) (Figure 3.18a). Unfortunately, for most leaf brain areas, LASSO fails to fit a model with very light regularization ($\lambda = 0.01$) to classify it against the rest of the brain in both the ST (mean train AUROC = 0.554) and the ABA (mean train AUROC = 0.593) (Figure 3.20a-d). The few leaf brain areas that are able to be classified from the rest of the brain using LASSO have a nearly identical performance in the one versus all case as in testing against all other leaf brain areas (Figure 3.18b; Figure 3.20a,b). At a higher regularization weight ($\lambda = 0.05$), most one versus all models fail to be trained (ST mean train AUROC = 0.501; ABA mean train AUROC = 0.502) (Figure 3.18b; Figure 3.20e-h). Failing to find potential marker genes using this approach with regularized LASSO, we turned to unregularized linear regression (i.e., $\lambda = 0$), with the hope to minimally find an identifying expression profile. Using linear regression, performance of models fit in the one versus all case correlates nearly perfectly with the average performance of the same model in one versus one. This nearly identical performance is true in both the ST (mean distance from

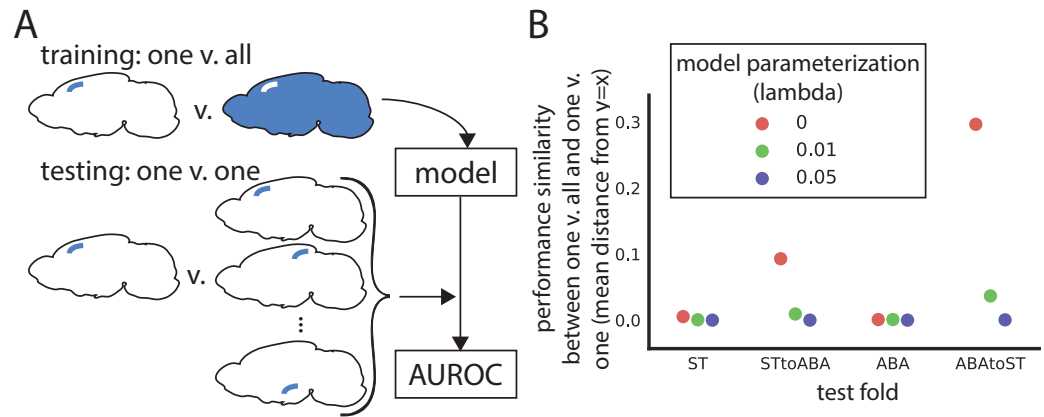


FIGURE 3.18: Identifying a unique gene expression profile for individual brain areas. (A) Schematic depicting training and testing schema for panels in this figure. Models are trained to classify one leaf brain area against the rest of the brain (one v. all) and then used to test classification of that brain area against all other leaf brain areas (one v. one) within and across dataset. (B) Performance similarity between one v. all and one v. one reported as the mean absolute value of distance from identity line for scatter plots of testing in one v. all against one v. one. Performance similarity shown for within ST, ST to ABA, within ABA, and ABA to ST across linear regression (red), LASSO (lambda = 0.01) (green), and LASSO (lambda = 0.05) (violet).

identity line = 0.005) and ABA datasets (mean distance from identity line = 0.001) (Figure 3.185b, Figure 3.19a, b) (see Section 3.3). This result demonstrates that within a dataset, we can find an identifying gene expression profile of a brain area that uniquely identifies it.

Since we could robustly identify a gene expression profile to identify a brain area within one dataset, we next asked if these profiles can generalize to the opposite dataset. Using the same models trained in one versus all in either the ST or ABA, we classified the same brain area against all other brain areas (one versus one) in the second dataset. The one versus all trained linear models (lambda = 0) do not generalize cross-dataset for either ABA to ST (mean distance from identity line = 0.296) or the reverse (ST to ABA mean distance from identity line = 0.093) (Figure 3.18b, Figure 3.19c, d). This lack of cross-dataset performance similarly holds for other parameterizations (Figure 3.18b; Figure 3.20c,d,g,h). The identifying gene expression profile of a leaf brain area is not generalizable to a new dataset that is not used in defining that profile. So, while we can uniquely identify a brain area using gene expression within one dataset, that identification profile does not extend to

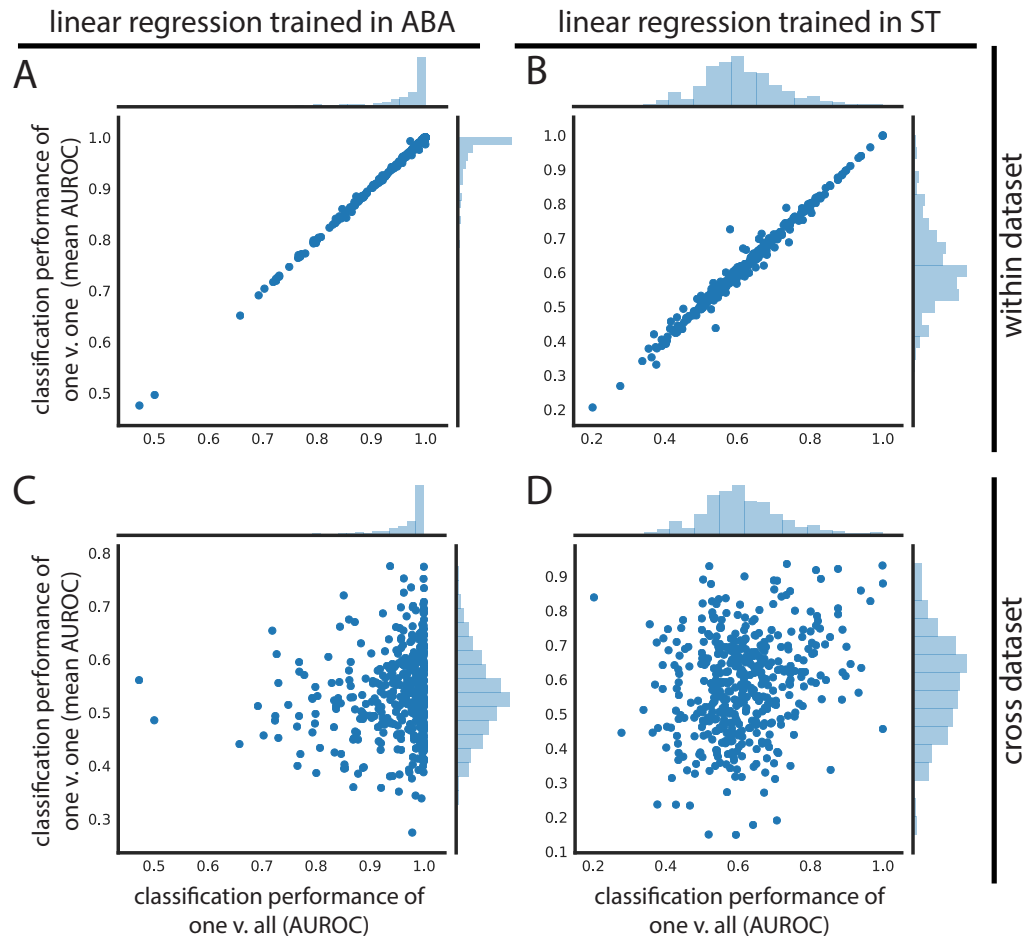


FIGURE 3.19: Leaf brain area expression profiles are identifiable within dataset, but do not generalize cross dataset. Linear regression one v. all test set performance (x-axis) versus average one v. one performance of the same model (y-axis) in (A) ABA and (B) ST. (C) Assessment of the same ABA one v. all linear regression model (x-axis) in one v. one classification in the ST dataset (y-axis). (D) Same as (C), but one v. all linear regression trained in ST (x-axis) and one v. one classification of these models in ABA (y-axis).

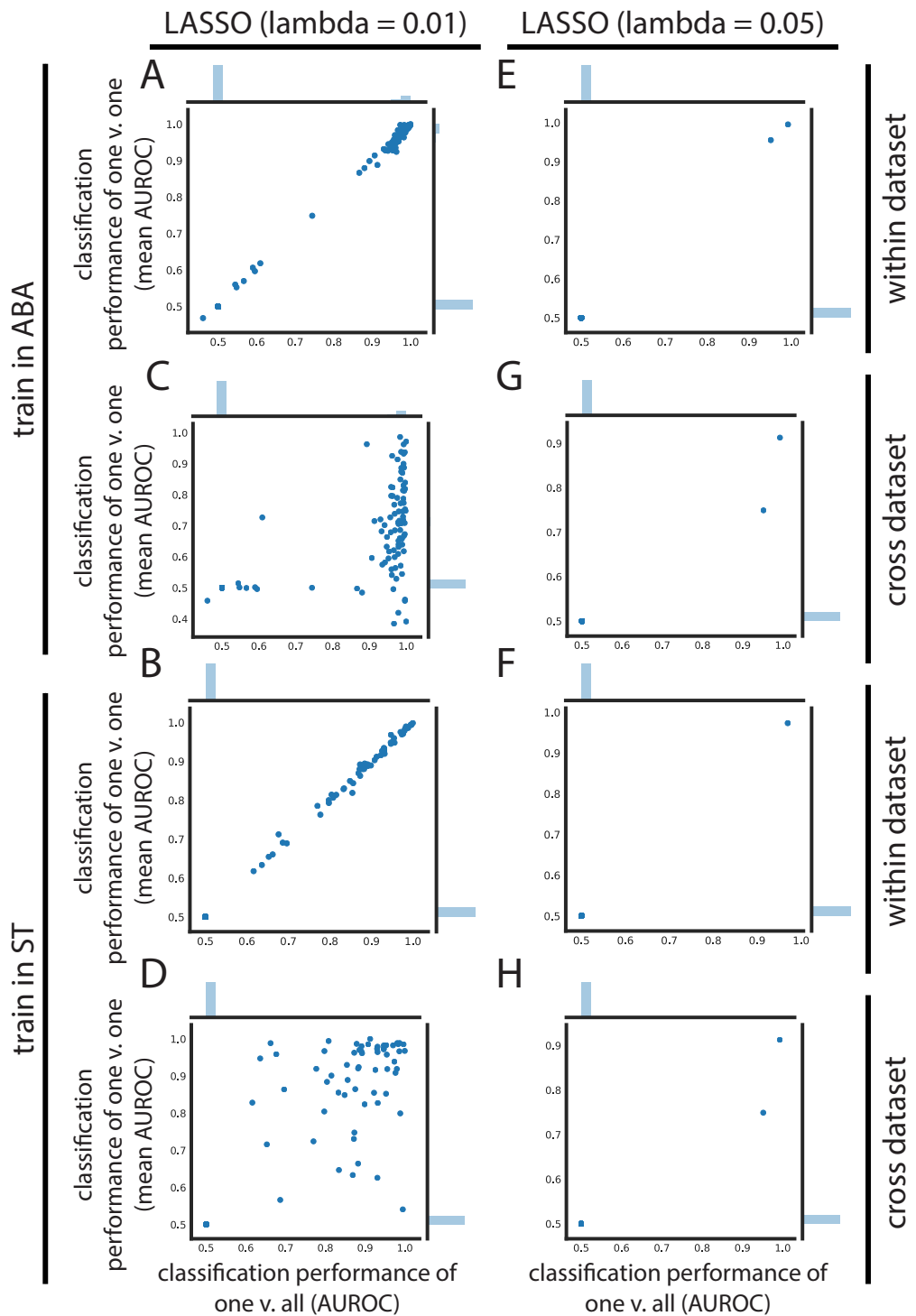


FIGURE 3.20: One v. all and one v. one analysis across various parameterizations. LASSO (lambda = 0.01) one v. all test set performance (x-axis) versus average one v. one performance (y-axis) of the same dataset (A) in ABA and (B) in ST. (C-D) Same as (A and B), but one v. one performance (y-axis) is accessed in the opposite dataset for (C) train one v. all in ABA and test one v. one in ST and (D) train one v. all in ST and test one v. one in ABA. (E-H) Same as (A-D) respectively, but using LASSO (lambda = 0.05).

the second dataset.

3.3 Methods

Spatial Transcriptomics Data (ST). Spatial Transcriptomics (ST) is an array-based approach where a tissue section is placed on a chip containing poly-T RNA probes that the mRNA transcripts present in the tissue can hybridize to (Ståhl et al., 2016). These probes tile the chip in 100 micron diameter spots and contain barcodes specific to that spot so that RNA sequencing reads can be mapped back to their original grid location. Note that the probe spots are not perfectly adjacent to each other, but have a center to center distance of $200\mu m$ (Ståhl et al., 2016).

Here, we used a previously published spatial gene expression dataset containing 75 coronal slices from one hemisphere of the adult mouse brain across three animals (Ortiz et al., 2020). The coronal slices were mapped to the Allen Mouse Brain Reference Atlas using a non-rigid transformation approach (Fürth et al., 2018). In total, this dataset contains 34,103 ST spots across 23,371 genes (Ortiz et al., 2020). On average, in one animal in this dataset, there were ~ 11 nuclei per spot.

Allen Brain Atlas in situ hybridization Data (ABA). The Allen Brain Atlas (ABA) adult mouse in situ hybridization (ISH) dataset consists of a transcriptome-wide assay of expression in inbred WT mice using single molecule ISH (Lein et al., 2007). To assay the whole transcriptome, many WT mouse brains were sliced into 25m thick slices containing 8 interlayered sets for subsequent single molecule hybridization or for staining to create the reference atlas. This results in a z resolution of $200\mu m$ for each gene. These independent image series are subsequently reconstructed to three dimensions and registered to the reference brain atlas in interlayered steps (Ng et al., 2007). There are 26,078 series, or experiments, across both coronal and sagittal planes with 19,942 unique genes represented. This suggests that a minimum of roughly 3,260 mice brains were used in this dataset, which does not include series that were unused or used for reference staining. These 3D registered reconstructions are then segmented to $200\mu m^3$ voxels with an associated brain area label. A rough estimate using a cell size of $10\mu m^2$ or $100\mu m^2$ would mean

80,000 or 8,000 cells per voxel. There are 159,326 voxels, with 62,529 mapping to the brain. Gene expression for each of the assayed genes was quantified in these voxels from the imaged data as energy values which is defined as the sum of expression pixel intensity divided by the sum of all pixels.

The quantified ISH energy values dataset was downloaded from the Allen Brain Atlas website through their API on March 12, 2019.

Allen Institute reference brain ontology, leaf brain areas, and path length. The Allen Institute reference brain atlas has organized brain areas into a hierarchy described by a tree data structure. Leaf brain areas are defined here as brain areas that constitute leaves on the ontology tree, i.e. they have no children. Leaf brain areas represent the most fine-scale parcellation of the brain. Using leaf brain areas circumvents the fact that the depth of the tree representing the hierarchical naming structure of brain areas in the ARA is not uniform. Path length refers to the number of steps required to go from one brain area to another in this tree.

Data filtering and train/test split. The ST data was pre-processed to remove ST spots mapping to ambiguous regions, fiber tracts, or ventricular systems and to remove genes that were expressed in less than 0.1% of samples. This left 30,780 ST spots, or samples, with 16,557 genes. For within ST analyses, this dataset was further filtered to 461 leaf brain areas that each had a minimum of 5 spots. In all analyses these spots are subsequently randomly split into train and test sets with a 50/50 split. The train/test split is random, but stratified for brain areas so that each fold has roughly 50% of the samples belonging to each brain area. N-fold (here, 2) cross validation was used and results are reported as a mean across folds.

Similarly, the ABA data was filtered for only voxels mapping to the reference brain and genes with expression in at least 0.1% of samples. This gives 26,008 series across 62,527 voxels also split as described for ST into 50/50 train and test folds. There are 4,972 genes that are assayed more than once across independent experimental series. Except for the analyses separating out the two planes of the ABA data (detailed below),

genes duplicated across series were averaged for each voxel for a total of 19,934 unique genes. For within ABA analyses, this dataset was further filtered to 560 leaf brain areas that each had a minimum of 5 voxels prior to the train/test split. As with ST, within ABA training and testing, n-fold (here, 2) cross validation was used and results are reported as a mean across folds.

For cross-dataset learning, both datasets were further filtered for 445 leaf brain areas that were represented with a minimum of 5 samples in each dataset. Genes were also filtered for those present in both datasets resulting in 14,299 overlapping genes between the two. This filtered subset was used for cross-dataset test set classification and matched within-dataset test set comparisons. For analyses separating out the two ABA planes, a similar mapping process was used to determine overlaps between each of the planes and the ST data. This resulted in 3737 overlapping genes across the same 445 leaf brain areas. Genes that were duplicated in the ABA dataset with independent imaging series within a plane were averaged.

Area under the receiver operating characteristic (AUROC), clustering using AUROC, and precision. The AUROC is typically thought of as calculating the area under the curve of true positive rate as a function of false positive rate. Here, the area under the empirical ROC curve is calculated analytically since it is both computationally tractable and accurate for a given sample. It is given by

$$AUROC = \sum_i^N \frac{Ranks_i}{N_{Pos} * N_{Neg}} - \frac{N_{Pos} + 1}{2 * N_{Neg}}$$

where ranks are the ranks of each positive label sorted by feature and N_{Pos} and N_{Neg} are the number of positive and negative labels respectively. This formula is based on the relationship between the Mann-Whitney U statistic and AUROC (Krzanowski and Hand, 2009; Hanley and McNeil, 1982; Mason and Graham, 2002). An AUROC of 0.5 indicates that the task being evaluated is performing at chance, while an AUROC of 1 indicates perfect performance. For within-dataset analysis (Figure 3.2, 3.8), any AUROCs of 0 were removed from downstream reporting of distributions and mean AUROCs. Note, this filtering

does not alter the reported means to the third decimal place. For within-dataset analyses, AUROC's are reported as the mean across 2-fold cross validation.

Clustering by AUROC is done by converting AUROC to a similarity metric by subtracting 0.5 to center the AUROC values at 0.5 and taking the absolute value. The rationale is that if a classification task performs with an AUROC of 0.5, the two classes are so similar that they are not distinguishable so they should be grouped closely.

Here, we calculate precision for the median performing brain area pair given by AUROC for within-dataset analysis. We use a threshold that includes all instances of one class, here, all instances of one brain area. Precision is calculated as:

$$precision = \frac{truepositives}{truepositives + falsepositives}$$

Note that the AUROC of median performing brain area pairs are calculated from the averaged AUROCs across 2 folds, while the reported precision is the average precision of all median brain area pairs from each fold independently because the reported median AUROC (from the fold averaged AUROCs) does not match to actual brain area pairs in either fold.

LASSO and penalty hyperparameter selection. Least absolute shrinkage and selection operator, or LASSO regression, uses a L1 penalty for fitting the linear regression model (Tibshirani, 1996). The cost function to minimize is given by:

$$cost\ function = \min_{\omega} \frac{1}{2n_{samples}} \|X\omega - y\|_2^2 + \lambda \|\omega\|_1$$

where X represents the matrix of feature values, y the target values, ω the coefficients, and λ the constant value with which to weight the regularization. The notation $\|\cdot\|_1$ represents the L1 norm. A small λ gives little regularization ($\lambda = 0$ is equivalent to regular linear regression). An L1 penalty minimizes the absolute value of coefficients, which has an effect of pushing many coefficients toward zero. This is beneficial for highly correlated data to find an optimal set of features among correlated genes, or features, to use for prediction.

In this manuscript, LASSO models are fit using coordinate descent according to the scikit-learn library (Pedregosa et al., 2011). Hyperparameter selection for the penalty weight λ is done through cross validation on a subset of brain areas that have sufficient sample size; we use a cutoff of having greater than 100 samples per brain area which resulted in 65 areas in ST and 139 areas in ABA. With this subset, we use the StratifiedShuffleSplit function from the scikit-learn library to create 3 folds with a test size of 20% within the 50% train set for each dataset (Pedregosa et al., 2011). These folds can overlap with each other, but are random and stratified by label. We next use these folds in the GridSearchCV function of scikit-learn to perform hyperparameter selection over λ values of 0.01, 0.05, 0.1, 0.2, 0.5, 0.9. Note, when returning the best performing hyperparameter or classification result as an AUROC, GridSearchCV returns the first of ties which can be misleading with tied performance across many hyperparameters (as is often the case here). In both ST and ABA, most pairwise brain area LASSO models perform best with the smallest given λ of 0.01 (Figure 3.21a,b). Since most brain areas lack the sample size to dynamically fit alpha, we chose a fixed λ value for all brain areas in our brain-wide analyses. Though there is not a clear trend, and keeping the ties or near ties in performance in mind, we use a λ of 0.1 for most of our analyses as larger lambda values tend to only show up for smaller brain areas. The hyperparameter λ used for each analysis is noted throughout the main text. We further perform hyperparameter selection in the cross-dataset case when the planes of ABA are separated out. Using 63 brain areas with greater than 100 samples across all three ‘datasets,’ we find that mean AUROCs across pairwise cross-dataset classification was comparable between the dynamically fitted λ and the fixed $\lambda = 0.1$ across brain area pairs (Figure 3.15c,d). For additional details on parameterization, see code scripts (repository availability below).

Linear Regression. Linear regression is implemented using scikit-learn with default parameters (Pedregosa et al., 2011). Normally, when there are more features than samples, linear regression is underdetermined. In the scikit-learn library, however, instead of returning linear regression as unsolvable, it returns the minimum Euclidean norm. (This is different from Ridge Regression where the L2 norm is incorporated in the cost function.)

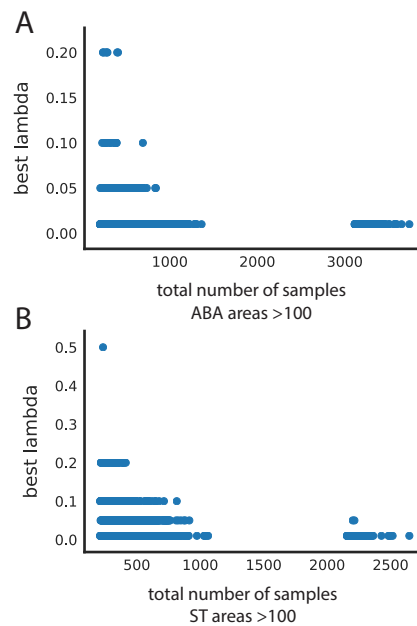


FIGURE 3.21: **Relationship between sample size and LASSO hyperparameter choice.** Plots showing the best lambda for LASSO when dynamically fit across various possible values as a function of brain area sample size in (A) ABA and (B) ST.

K-nearest neighbors algorithm (k-NN). For applications of k-nearest neighbors (k-NN) we used the scikit-learn implementation with default hyperparameters: $k = 5$, `weights = 'uniform'`, `algorithm = 'auto'` (Pedregosa et al., 2011). Similar to LASSO, 50% of the data was used as the train set, calculating performance of classification on the other 50% held out test set. As an output, k-NN gives a 1D vector with the length equal to the number of samples in the test set. This vector contains a predicted brain area label for each test set sample based on the most highly represented class among each test set sample's k closest neighbors in expression space. To compare this classification result to our other classification approaches, we separated out the 1D vector into a 2D binary matrix with a column for each brain area and rows of the same length as the 1D vector representing samples. Each time a sample is predicted as being a particular brain area, the corresponding row and column are marked with a 1. This matrix is then used to calculate an AUROC for predicting each brain area, or column. The mean AUROC of these brain areas is reported in the manuscript.

Batch correction using pyComBat. For batch correction, we use pyComBat, a

recent python based implementation of ComBat (Behdenna et al., 2020; Leek et al., 2012; Johnson, Li, and Rabinovic, 2007). Prior to batch correcting, we normalize each dataset independently using z-scoring. We then run pyComBat on the two datasets combined treating each dataset as a batch. We do not include covariates. Corrected data is then parsed into the two datasets from the combined matrix for subsequent cross-dataset LASSO analysis.

Assessing dimensionality of data using Principal Component Analysis (PCA).

Principal Component Analysis (PCA) as implemented in scikit-learn (Pedregosa et al., 2011), was used to determine the dimensionality of both datasets. PCA was applied to the genes, or features, for each leaf brain area separately in the ABA and ST datasets. The total number of components to use for dimensionality reduction was set to be equal to the number of samples in each area. ABA areas were down-sampled to have the same number of samples as the corresponding brain area in ST. There are 77 brain areas that exceptionally have fewer samples in ABA than ST, so when down-sampling ABA for these 77 areas, the original sample size was used. Dimensionality of the brain areas were then accessed as the number of PCs needed to explain at least 80% of the variance.

Differential expression and correlation-based feature selection (CFS). Differential expression (DE) in genes is assayed using Mann-Whitney U (MWU). Resulting p-values are not corrected for multiple hypothesis testing since p-values are only used to threshold for very extreme DE genes across brain area comparisons. The un-corrected p-values themselves are not reported as a measure for significant DE.

Correlation-based feature selection (CFS) is a feature selection technique that explicitly picks uncorrelated features (Hall, 1999). Here, a greedy approach to CFS was implemented. The algorithm first chooses a random seed or gene within the top 500 differentially expressed genes. The next gene is then chosen as the lowest correlated gene to the first one and kept if the set AUROC improves. Subsequent genes are chosen as the least correlated on average to the genes already in the feature set. The algorithm stops once the AUROC is no longer improving. The final set of genes chosen using CFS are then aggregated by equally by averaging the values of all chosen genes for each sample. Here,

in particular, these feature sets fell in the range of 1-29 genes with a median of 2 genes in the ABA and the range of 1-47 genes with a median of 4 genes in the ST (Figure 3.13a,b). For more details on exact implementation see code scripts (repository available below).

For the cross-dataset analysis, when un-specified, 100 feature sets were chosen using this approach and the single best performing feature set was then evaluated in both the within-dataset test set and the cross-dataset test set. When indicated accordingly, the 100 CFS feature sets were averaged instead of reporting the performance of the best set alone.

Euclidean distance between two brain areas. In addition to brain area labels, the ABA dataset contains x, y, z coordinates for each voxel in the ARA space. So, physical distance between two brain areas is calculated as the Euclidean distance between the two closest voxels where each voxel belongs to one or the other brain area. Due to the symmetry of brain hemispheres, distance was only calculated in one hemisphere by filtering for voxels with a z-coordinate less than 30. This z-coordinate was visually determined to be the midline of the brain based on 3-D visualization of the voxel coordinates. Euclidean distances between brain areas calculated in this manner were used for both the ST and ABA datasets since both are registered to the ARA.

Mean distance from identity line. To assess the replicability of models trained in one brain area versus the rest of the brain (one versus all) in classifying that same brain area against all the others (one versus one), the mean absolute Euclidean distance of a scatter plot of those two values from the identity line was calculated. This was done to assess how similar the values are in the one versus all case are to the one versus one case for each pair of brain areas. Correlation was found to be lacking because it could yield high correlations when the one versus all and one versus one values were quite different for a given point.

Code. All code used for the analyses described in this manuscript was written in Python 3.7 with supporting packages: jupyterlab 1.0.9, h5py 2.9.0, numpy 1.16.4, scipy 1.3.1, pandas 0.25.0, scikit-learn 0.21.2, matplotlib 3.1.0, and seaborn 0.9.0. All Jupyter notebooks and scripts are available on GitHub at: www.github.com/shainalu/

spatial_rep.

3.4 Conclusions

Across disciplines, benchmarking studies have helped to advance their respective fields and set standards for future research (Shi et al., 2006; Su et al., 2014; “Meta-analysis in basic biology” 2016; Robinson and Vitek, 2019). Neuroscience is no exception. Given the complexity of the brain, however, the addition of multimodal information is especially desirable. Studying the brain from a variety of perspectives, across data modalities, technology platforms, and experiments, can give a complete, composite understanding of its biology (Lein, Borm, and Linnarsson, 2017). Coupled with the stereotyped sub-structure and transcriptional heterogeneity, the adult mammalian brain is the ideal model system to assess new spatial gene expression technologies. With this in mind, we linked two modalities to ask: can we capture canonical, anatomically-defined brain areas from the Allen Reference Atlas using spatial gene expression alone? And, how well does this replicate across two transcriptomic datasets collected using different platforms?

Principally, we showed that ARA brain labels are classifiable using only gene expression, but highlighted a lack of generalizability across spatial transcriptomic datasets. Within datasets, we are able to distinguish brain areas from each other with high performance. We were further able to uniquely identify a brain area within dataset; training on one brain area against the entire rest of the brain generalizes to testing on that same brain area against all other leaf brain areas. Notably, within-dataset performance was on average higher in the ABA than the ST, which led to a lack of cross-dataset generalizability when training in the ABA and testing in ST; this phenomenon was not present in reverse. However, in both cases, there is an observed trend linking an increase in mean cross-dataset performance with increased semantic distance in the ARA brain area label organization. There was no link in performance when compared to physical distance in the brain suggesting ARA labels are meaningful in expression space and we are not simply detecting spatial differences in gene expression.

It is important to point out that our benchmarking study at its core only involves two independent datasets, although both are extraordinary in scope. Limited to two datasets, it is impossible to tell whether one dataset or the other is closer to representing the ground truth. Further, the ABA and ST datasets use two fundamentally different techniques: the ABA data reports average pixel intensity from in situ hybridization and the ST approach is an RNA capture technique followed by sequencing that reports read counts. In addition, the spatial resolution varies between the two datasets. Each sample in the ST is further apart within a plane due to a 200 μ m center to center distance for probe spots (Ståhl et al., 2016) in comparison to samples in the ABA ISH with 200 μ m³ voxels that tile adjacently (Lein et al., 2007). Conversely, the ABA ISH has a lower Z resolution, or larger gap between slices (200 μ m) when compared to the ST (median slicing period of 100 μ m) (Ortiz et al., 2020). Further, after the collection of the raw expression data, each of the two datasets also undergoes a unique registration step to the ARA. The ABA uses an iterative approach that involves registration of the 3D brain volume with interspersed smoothing steps (Ng et al., 2007), while the ST dataset is registered on a slice by slice basis to the nearest representative 2-D ARA slice using anatomical landmarks (Fürth et al., 2018). Beyond registration, there are additional concerns about the stability of the ARA brain area labels since there are inconsistencies with other brain atlases and even across versions of the ARA (Azimi et al., 2017; Chon et al., 2019; Wang et al., 2020). Brain atlases are an imperfect formalism of brain substructure, but are the best systematic representation to test spatial gene expression by biological areas.

Past these technical differences, it is also possible that the lack of strong cross-dataset generalization could represent true biological brain to brain variability of individual mice. Slight changes in cellular composition between individuals near borders of brain areas could be responsible for the differences between these areas. Zooming out, the results of this manuscript have many potential implications for neuroscience. First, we observed that brain regions are comprehensively better defined by a combination of many genes as opposed to individual markers. Traditionally, however, the description and/or subsequent

experimental identification of brain areas by gene expression, often via specific populations of neurons, usually depended on one or two marker genes (Grange and Mitra, 2012; Huang, 2014). The choice to use single marker genes is usually one of practicality, but as spatial experimental and computational analysis techniques continue to improve, the possibility of using better resolved brain areas defined by a multi-gene expression profile is within reach (Ortiz et al., 2020). This in turn could inform downstream experimental identification of brain areas. Secondly, our results show that quantitative exploration of the existence of brain regions is timely, just as single cell data has made quantitative exploration of cell type definitions timely. The definition of brain regions is necessary to study the brain, but existing parcellations are almost certainly not sufficient; it is easy to overfit to rigidly defined areas, glossing over individual differences. This highlights the need for continued development of approaches for validating and integrating spatial data from multiple sources. The ST data is a technology which is timely to integrate with single cell data and valuable to validate, refine, and discover an iterative neuroanatomy that grows with new data types and sources (Toga et al., 2006). Finally, an iterative, multi-modal definition of brain areas could aid in the research of cross-species comparisons and disease phenotypes, where the mapping of neuroanatomical landmarks is potentially complex.

As the types and prevalence of spatial gene expression approaches continue to increase (Asp, Bergenstråhle, and Lundeberg, 2020) (see Section 2.2), whole-brain spatial gene expression datasets will surely follow. By continuing to integrate these emerging datasets, we will be able to perform more robust meta-analyses, giving us a deeper understanding of both spatial gene expression with respect to ARA labels and the replicability of spatial technologies in general. An added benefit of the continued incorporation of additional datasets, is that at some point, differences in experimental platforms and registration approaches will only contribute to the robustness of any biological claims. We believe continued meta-analysis of spatial gene expression in the adult mouse brain and other biological systems is an important route toward integration of distinct data types – location and expression – to form the beginnings of a robust, multi-modal understanding of the mammalian brain and other systems.

3.5 Future directions

Looking forward, there are a handful of immediate steps to build on this work. Firstly, for the cross-dataset assessment, we could additionally train a model combining samples from both datasets and then compare performance to held out test data both separated by dataset and combined. Likely, performance of such a model would outperform both individual models in generalization since it sees examples of data from both datasets in training. Further, given the differences in the detection of genes between ISH and ST methods (see Section 2.2.5), it would be interesting to subset the genes used in our models. For example, restricting the analyses to highly expressed genes could change the performance of classifying datasets either within or across dataset. Lowly expressed genes tend to be better detected with ISH than sequencing-based methods, and perhaps the very high performance seen in ABA that fails to generalize to ST could be driven by these lower expressed genes that are not well captured in the ST.

Stepping back, registration and quantification of signal from experimental brain slices could be improved both internally in the ABA and externally. Regarding the ABA, we suspect that some of the high performance in classification could be driven by overfitting during iterative registration. Given the scope (whole-brain, whole-transcriptome), re-quantification of the ABA was well beyond the scope of this chapter. (It was additionally unclear that a second full re-quantification would yield a better result.) Perhaps, however, some careful consideration of the original registration and removal of some of the iterative steps could improve registration. For example, the ABA ISH dataset is registered in 3-D (Ng et al., 2007), perhaps registration in 2-D as is done with the 'external' ST dataset could make the performance of the two more comparable. Although not on the whole-transcriptome, whole-brain scale, we explore some re-quantification of the ABA in the next chapter (see Section 4.2) when comparing it to BARseq data. Externally, registration could be more robust as well. While the tools of registration are well-supported and easy to use (Fürth et al., 2018), the selection of comparable slices in the ABA to align

new experimental slices to is often done by eye and somewhat arbitrarily. Additional normalization of selecting comparable slices along the axis of slicing could make registration more uniform.

Independent of registration, brain regions defined by cytoarchitecture as in the ARA is not completely compatible with gene expression. Each of these brain regions consists of a variety of cells with different gene expression programs- commonly referred to as cell type heterogeneity. Furthermore some cells, particularly non-neuronal cells are present throughout the brain. The expression signature of these cell types could make it harder to distinguish between brain areas based on expression alone. Integration of single-cell data with this spatially-resolved data in the mouse brain could answer questions about the contribution of differential cell type composition to spatial patterning of expression. In Section 4.3, we take a stab at doing exactly this in only the primary motor cortex, interacting with single-cell data at the level of using cell-type marker genes previously defined from scRNA-seq. A more comprehensive and thorough integration of spatial and single-cell would offer much in elucidating the underlying biology.

Another obvious direction of this work is the creation of a spatial gene marker database. With additional datasets, we could define robust markers of spatial expression that could be subsequently used in experimental neuroscience to identify the corresponding regions. While such a database would provide a valuable resource, given the multi-gene profiles of spatial expression as discussed above, such markers may be hard to identify. Finally, as discussed in the previous section (see Section 3.4), the ultimate follow-up of this work is the continual integration and meta-analysis of additional spatially-resolved datasets that are similarly whole-transcriptome and whole-brain in scope. As spatially-resolved techniques continue to be developed and democratized (see Section 2.2), the ease of collection and availability of such datasets should follow.

Chapter 4

Vignettes of spatial transcriptomics in neuroscience

4.1 Other approaches to spatially-resolved transcriptomics in neuroscience

There are many facets of spatially-resolved transcriptomics as it relates to neuroscience. In the previous chapter (Chapter 3), we showed that despite the ability to capture canonical brain area labels with expression in each dataset individually, there was a lack of replicability when going cross-dataset. We posit that much of this discrepancy could be attributed to the registration of the ABA dataset and trying to use it in a global sense, which is not the main intention of the data. In this chapter, I explore various aspects of spatial expression in neuroscience on a smaller-scale by focusing in on specific genes and/or brain areas. We first use the raw ABA ISH images to benchmark BARseq (see Section 2.2.3 and Section 2.4.3) detected expression in the auditory cortex (Section 4.2). Next we probe the relationship between the cell-type markers defined using single-cell data and markers of spatial expression to ask if spatial patterning of expression is driven by cell type composition in the visual cortex (Section 4.3). Finally, we implement image analysis methods to read out sequences from BARISTAseq (see Section 2.2.3) *in situ* sequencing images (Section 4.4).

4.2 Replicability of BARseq2 with re-quantified ABA ISH

The results reported in this section are adapted from the Nature Neuroscience Technical Report titled "Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections," which was authored jointly by Yu-Chi Sun, Xiaoyin Chen, Stephan Fischer, Shaina Lu, Huiqing Zhan, Jesse Gillis, and Anthony Zador. The full published text is available in Appendix D and Sun et al., 2021. Here, I focus on my contribution to this manuscript, the comparison of BARseq data with the Allen Brain Atlas, and relevant context.

4.2.1 Brief introduction to cadherins

Cadherins, short for calcium-dependent adhesion, are a large group of transmembrane proteins responsible for cell-cell adhesion. Their presence is wide-spread, including in the brain. Cadherins are thought to have a role in development, including cortical development and guiding projections (Hayano et al., 2014; Friedman et al., 2015). Further cadherins were previously reported to show differential expression between sub-types of GABAergic neurons (Paul et al., 2017), suggesting potential spatial patterning of expression dependent on differential cell type composition. Indeed, cadherins were previously shown to have differential spatial expression across layers of the cortex (Matsunaga et al., 2015). These properties make cadherins an interesting gene family to study in spatial resolution in conjunction with projections using BARseq2. Briefly, BARseq2 is a multi-modal neuroscience approach which combines *in situ* sequencing with the sequencing of barcoded projections (see Section 2.4.3).

4.2.2 Brief introduction to the auditory cortex

The auditory cortex is a region of brain involved in processing auditory inputs. The auditory cortex is usually sub-divided into three or four sections. In the ARA there are four sub-divisions: the dorsal auditory area (AUDd), primary auditory area (AUDp), posterior auditory area (AUDpo), and the ventral auditory area (AUDv) (Lein et al., 2007). The AUDp is the primary recipient of the corresponding thalamic area involved in auditory

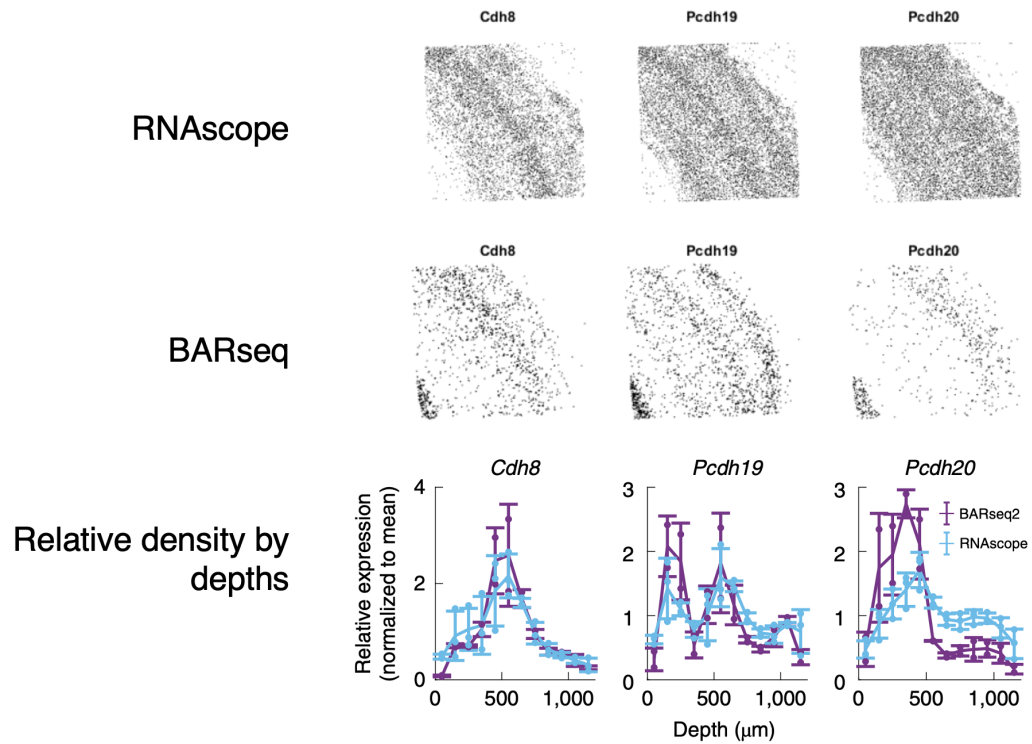


FIGURE 4.1: **Comparison of Cdh8, Pcdh19, and Pcdh20 BARseq expression with RNAscope in primary auditory cortex.** Top: RNAscope *in situ* hybridization images; Middle: BARseq2 *in situ* sequencing images; Bottom: Quantification of RNAscope detected expression (light blue) with BARseq2 detected expression (purple). Lines indicate means, error bars indicate s.d. values, and dots show individual data points. $n = 2$ slices for BARseq2 and $n = 3$ slices for RNAscope. *Images and plots courtesy of Yu-Chi Sun and Xiaoyin Chen; Bottom is published in Sun et al., 2021.*

processing and thus has matching tonotopic map to this area (Purves et al., 2001b). Similar to other cortical areas and their respective sensory tissue, this map is a representational map of its sensory tissue- the cochlea. This map is slightly different in the AUDp than other cortical areas since the cochlea already has a tonotopical arrangement that is mapped in the auditory cortex. In mice, the auditory cortex is located on the sides of the brain. This can make it harder to perform experiments relative to other cortical areas and perhaps limits the research done on this area relative to extremely well-studied cortical areas closer to the top of the brain where experiments are more easily done (e.g. visual cortex).

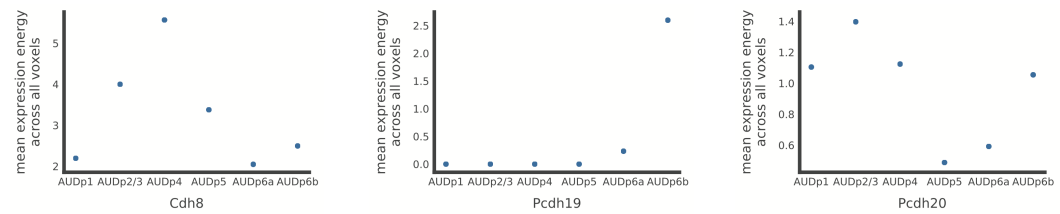


FIGURE 4.2: **Expression energy from Allen of Cdh8, Pcdh19, and Pcdh20 across layers of the primary auditory cortex.** Dots represent the mean expression energy from all 200m^3 voxels assigned to the given layer of the primary auditory cortex.

4.2.3 Re-quantification of cadherins from raw in situ hybridization images

Without definitive evidence, the work in Chapter 3, suggests that despite the high signal of ISH in the ABA, some of the expected spatial patterning of expression may be lost or distorted in the subsequent registration and quantification done by the Allen Institute. Focusing on specific genes, BARseq2 was applied to 20 cadherins plus 3 marker genes in the primary auditory cortex in mice. RNAscope was used to validate endogenous expression from BARseq2 for 3 genes: Cdh8, Pcdh19, and Pcdh20 (Figure 4.1). Comparing Allen quantification with BARseq2 and RNAscope, we indeed observed that expression patterns that were evident in the raw images of the ABA themselves, BARseq2, and RNAscope were not reflected in the quantified values (Figure 4.1 and 4.2) (see Section 4.2.4). For example, expression of Pcdh19, which was only available in sagittal slices wholly did not match BARseq2 or RNAscope expression or its own ISH image (Figure 4.2).

After further investigating the sagittal and coronal ISH images in genes that were assayed in both planes, we determined that using only the genes assayed in coronal sections would be the best approach as coronal ISH experiments in the ABA were generally of higher quality and collected later in the creation of the database. All cadherins detected with BARseq2 that had coronal images (to also match the slicing direction for BARseq2) in the ABA were validated against the ABA ISH data. 16 genes were validated in this manner, which only leaves out 4 BARseq2-assayed genes in the cadherin family that were not validated against ABA. In addition, we shifted from using the Allen quantification of the ABA to re-quantification from the ISH images themselves (see Section 4.2.4). In this manner, we found that expression was well replicated between BARseq2 cadherins and the

ABA (Figure 4.3). Laminar expression was better correlated between ABA and BARseq2 when compared to a random shuffling of expression across positions (Figure 4.4).

In summary, for the subset of genes in the cadherin family explored here, spatially-resolved expression across different technologies (see Section 2.2) replicate with each other. Technologies spanning ISS (BARseq2), modern ISH (RNAscope), and older ISH approaches (ABA) detected similar laminar patterns of expression in the mouse cortex. Crucially ABA ISH requires re-quantification from the raw-images on a gene-by-gene basis to better match the processing done with the other two approaches. The prior global registration and quantification of ABA done by the Allen Institute themselves led to some confusing outputs here. The replicability, if even on a small scale (handful of genes) found here is a hopeful harbinger of the potential of various spatially-resolved transcriptomics approaches to assay meaningful biology (see Section 2.5).

4.2.4 Methods

Plotting Allen quantification of ABA. Allen quantification is reported as an "expression energy" which is defined as

$$\text{expression energy} = \text{expression intensity} * \text{expression density}$$

where

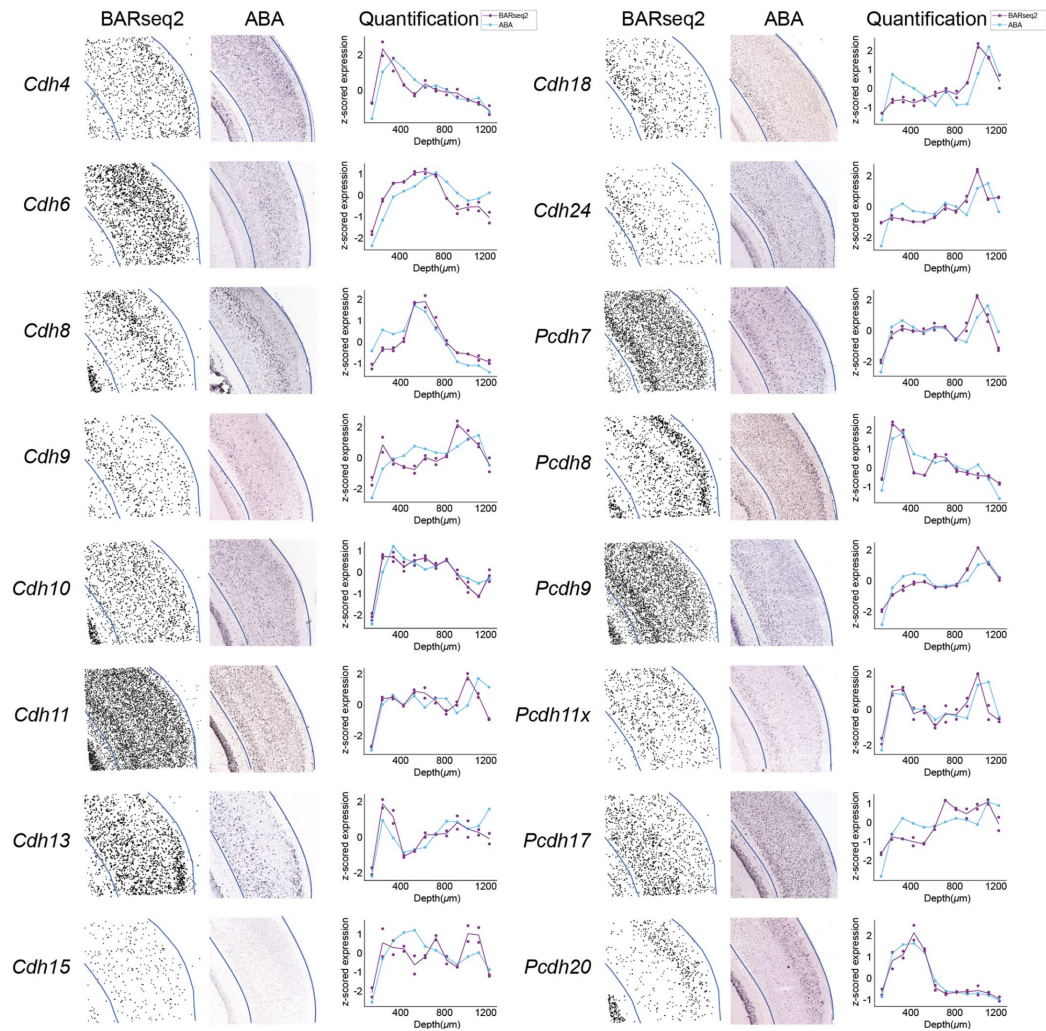
$$\text{expression intensity} = \text{sum of expressing pixel intensity} / \text{sum of expressing pixels}$$

and

$$\text{expression density} = \text{sum of expressing pixels} / \text{sum of all pixels in division}$$

(Lein et al., 2007). When reporting expression energy, here we average across all series (sagittal and coronal, when available) that are available in the ABA. Accessing data, filtering, and pre-processing for working with the voxel quantification of the ABA is described in Section 3.3.

FIGURE 4.3: **Comparison between BARseq2 and ABA *in situ* expression atlas for cadherins.** Gene expression patterns in auditory cortex identified by BARseq2 are plotted next to *in situ* hybridization images of the same genes in Allen gene expression atlas (ABA) and the quantified laminar distribution of the gene in both datasets. Only genes that had coronal images in the Allen gene expression atlas are shown. Blue lines indicate the boundaries of the cortex in both BARseq2 and ABA images. In the laminar distribution plots, dots represent values from two BARseq2 samples (purple) and one ABA sample (blue) per gene. Lines indicate means across samples. *This plot is published in Sun et al., 2021.*



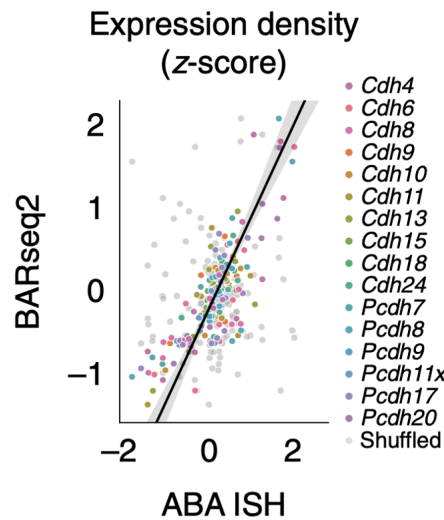


FIGURE 4.4: **Summary of relative gene expression observed using BARseq2 and in Allen gene expression atlas.** Each dot represents the expression of a gene in a 100- μ m bin in laminar depth. Gray dots indicate the correlation between data randomized across laminar positions. A linear fit and 95% confidence intervals are shown by the diagonal line and the shaded area. $n = 2$ slices for BARseq2 and $n = 1$ slice for ABA ISH. *This plot is published in Sun et al., 2021.*

Re-quantifying ABA ISH from images. Re-quantification of ABA ISH was done in multiple steps. First a coronal slice of the ABA was chosen to match the z-resolution of the slice assayed with BARseq2. Next, the auditory cortex ROI was selected in matlab. Using a custom script, vertices of the ROI are selected manually with a mouse click in the composite image. Additionally top and bottom lines are drawn around the ROI. Coordinates for the top and bottom lines and pixel (x,y) coordinates of the ROI are saved as csv files and exported to python. In python the ROI pixels are visualized with the original ISH images to check that the ROI is as expected. Then in each imaging channel, a simple cutoff is chosen based on graphs showing pixel intensities of each channel. This cutoff is used to 'binarize' the image and determine an expressing pixel from a background pixel. Since the ROI pixels and the top/bottom boundary of the ROI are collected independently, some pixels near the border may be included in the ROI, but not within the top/bottom boundaries. To filter out these pixels beyond the top/bottom boundaries, we calculate the angle between each ROI pixel and the top and bottom boundaries. Pixels within the boundaries will have angles near 180° , while pixels beyond the boundaries will have an angle less than 90° and

near 0° . We filter out ROI pixels near 0 to remove those beyond the top/bottom boundaries. Next the depth of each pixel in the cortex is normalized using a min-max normalization to 0 to 1 and multiplied by the max cortical depth (around 1200) to re-scale. Finally, the ISH quantification is reported as the (sum of all pixel values where each pixel value is averaged over the three channels)/(total number of pixels). This value is z-scored before comparing to BARseq2 quantification which is also z-scored.

4.3 Probing the relationship between spatial and single cell data in the primary motor cortex

The results reported in this section are done in collaboration with Stephan Fischer and my Partners for the Future mentee Elyse Schetty.

4.3.1 Brief introduction to cell-type markers defined from scRNA-seq

Cell types are defined by a variety of characteristics (Clevers et al., 2017; Arendt et al., 2016). This is especially true in the brain where cells have a variety of phenotypes including physiology and morphology. However, the recent proliferation of single-cell transcriptomics technologies has renewed interest in a transcriptomic-based definition of cell types (Tasic et al., 2018; Zeisel et al., 2018; Fischer and Gillis, 2021). In this paradigm, marker genes are simply those that are highly expressed in a specific cell population and lowly expressed in other cells. Cell type markers can be defined by clustering scRNA-seq cells based on their transcriptomes and then identifying DE genes. This can be done on single datasets or across multiple datasets for robust metamarkers. Here we use metamarkers defined from 7 datasets (Fischer and Gillis, 2021) to probe the relationship between spatial patterning of expression and the distribution of various cell types in space.

4.3.2 Brief introduction to the motor cortex

The motor cortex is a region of the brain involved in initiating movements of the body. In primates, including humans, electrical stimulation of different parts of the motor cortex have been shown to elicit muscle contractions on different parts of the body (Purves

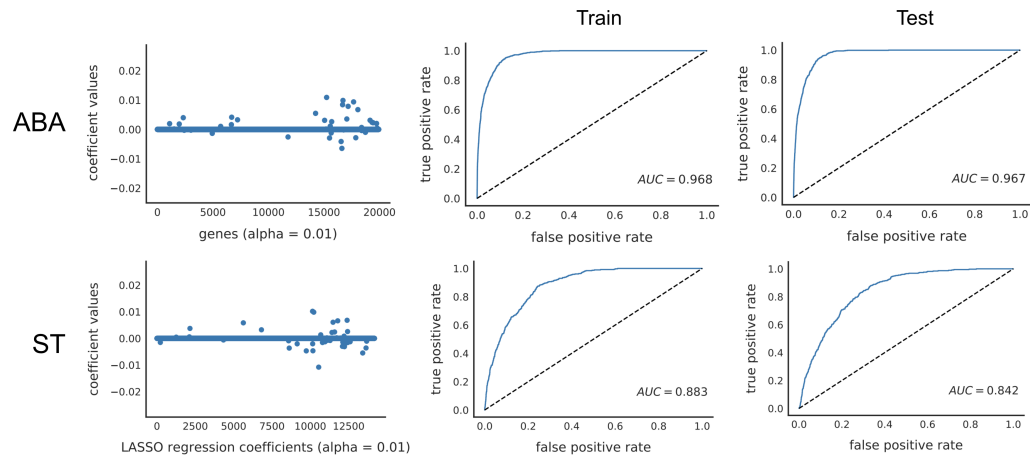


FIGURE 4.5: **The primary motor cortex is classifiable using gene expression.** Left: coefficients of a trained LASSO model ($\alpha = 0.01$). Note that the line at 0 is not a line but many genes with a coefficient of 0. Middle: AUROC curves of performance in the train set. Right: AUROC curves of performance in the test set. For all three columns, top row is the ABA data and bottom row is ST.

et al., 2001a). A complete topographical map of the motor cortex is known. Matching disproportionate representation of sensory areas in the somatosensory cortex, the motor cortex also has large areas mapping to areas with high fine motor cortex (e.g. hands) when compared to areas with more gross motor control (e.g. trunk of body). The motor cortex is of particular interest to us, since it was the focus of the Brain Initiative Cell Census Network (BICCN). BICCN has produced many molecular datasets focused on the motor cortex.

4.3.3 In the primary motor cortex, are spatial expression patterns merely capturing cell type composition differences across the brain?

We specifically chose to do a targeted analyses of the primary motor cortex (MOp) given the abundance of single cell RNA-sequencing (scRNA-seq) data through the BICCN collaboration (Yao et al., 2020; Network (BICCN) et al., 2020). We asked (1) if there are spatial markers of gene expression that replicated across the ST and ABA datasets (described in 3) and (2) whether these markers were overlapping with known cell-type markers defined from the BICCN data (Fischer and Gillis, 2021). The MOp could be learned using gene expression (Figure 4.5). Models trained in the ST generalized to the ABA, but not

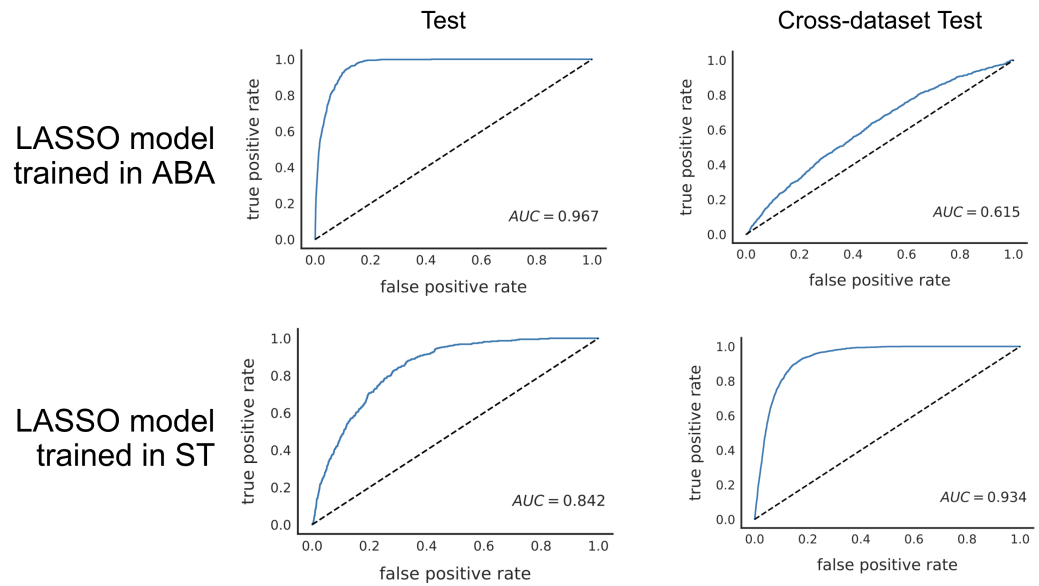


FIGURE 4.6: Cross-dataset performance of identifying the primary motor cortex Left: AUROC curves of performance of classification in the within-dataset held-out test set. Right: AUROC curves of performance of classification in the opposite dataset. For both columns, top is LASSO models trained in ABA and bottom is LASSO models trained in ST.

the reverse (Figure 4.6). This follows more general trends observed when looking at all brain areas (see Chapter 3, Figure 3.10). Finally, we compared spatial markers of the MOP with cell type markers to determine if there was a significant overlap between the two. For markers of excitatory, inhibitory, or non-neuronal cell-types, there was no evidence of enriched overlap between the genes chosen as markers of the MOP and any of these cell-types (Figure 4.7) (see Section 4.3.4). In corollary, there was no evidence of enriched overlap with markers of GABAergic and excitatory sub-types (Figure 4.8).

While there is no evidence of spatial patterning being driven by compositional cell-type effects here, these results are preliminary and use spatial markers defined using only 2 spatially-resolved datasets and are applied only to one brain area. Expression across different cortical areas is similar and differences in expression in the cortex is primarily driven by cortical layers and not areas (Codeluppi et al., 2018). Looking at the MOP as a whole relative to the rest of the brain, including other cortical areas, is perhaps less likely to yield useful spatial markers that relate to broad cell-types. Perhaps either zooming in and looking at specific layers of the MOP or zooming out and looking at the cortex as a

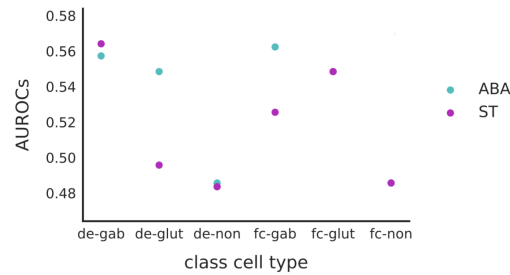


FIGURE 4.7: **Motor cortex gene markers are not enriched for broad cell types.** AUROC (y-axis) of comparison of DE genes with high expression in MOP to broad brain cell type markers does not show a significant overlap.

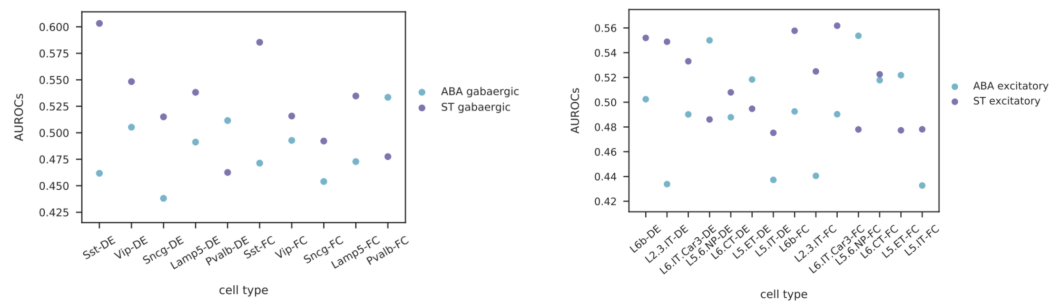


FIGURE 4.8: **Motor cortex gene markers are not enriched for GABAergic and excitatory sub-types.** AUROC (y-axis) of comparison of DE genes with high expression in MOP to sub-type markers (x-axis) of GABAergic (left) and excitatory (right) neurons does not show a significant overlap.

whole would identify spatial markers that are driven by cell-type differences. Additionally, perhaps other brain areas (e.g. sub-cortical areas) are more likely to have spatial patterning driven by cell-type distributions. Future experiments should probe these possibilities as robust cell-type-DE markers defined from the whole brain or other areas become available.

4.3.4 Methods

Classification of MOP vs. the rest of the brain and cross-dataset learning using LASSO. Classification of the MOP using LASSO and cross-dataset learning in the MOP are all done according to the methods described in Section 3.3. The only difference is that the data was subsetted to compare the MOP against the rest of the brain instead of pairwise comparisons of all leaf brain areas as described in Section 3.3. Additionally, LASSO hyperparameter selection here was done manually after inspecting the test set results across

various values.

Enrichment between MOp spatial markers and cell-type markers. To find markers of the MOp, we performed a one-tailed Mann-Whitney U test where the alternative hypothesis is:

$$H_A = MOp\ expression > not\ MOp\ expression$$

After this DE calculation, we then rank sorted the genes by p-value. Genes higher expressed in the MOp fell towards the top of this list. The single cell metamarkers were similarly ranked within each cell type so that the best markers were at the top of the list. Then to determine if there is overlap in the top of the list, an AUROC is calculated between these two lists. If the AUROC is near 1 it would indicate a high amount of overlap between the two lists toward the top of the lists. This would mean that there was enrichment between the single-cell cell type markers and the spatial gene markers. For each cell type (broad and subtypes), two different spatial metamarker lists were used that were defined either according to fold change or DE (Fischer and Gillis, 2021).

4.4 Reading out sequences for *in situ* sequencing

The results reported in this section are done in collaboration with Xiaoyin Chen.

In ISS protocols, raw data captured in microscopy images must be processed into their final sequenced read-outs (see Section 2.2.3 and 2.3). As previously described there are a variety of ways to do this base-calling. Here we describe one such way optimized for the read-out of barcodes (as opposed to endogenous mRNA transcripts) for use with spatial sequencing of MAPseq barcodes (see Section 2.4.3).

Prior ISS approaches relied on the assaying of endogenous mRNA transcripts by mapping potential reads to a reference genome as part of their base-calling protocol (Figure 4.9) (Lee et al., 2015). This alignment step was used to cluster neighboring pixels of the sequencing image that had the same reads. Adapting ISS to read out random oligonucleotide barcodes for MAPseq meant that there is no reference genome available. Since it is costly and difficult to get the sequencing depth needed to represent a random barcode

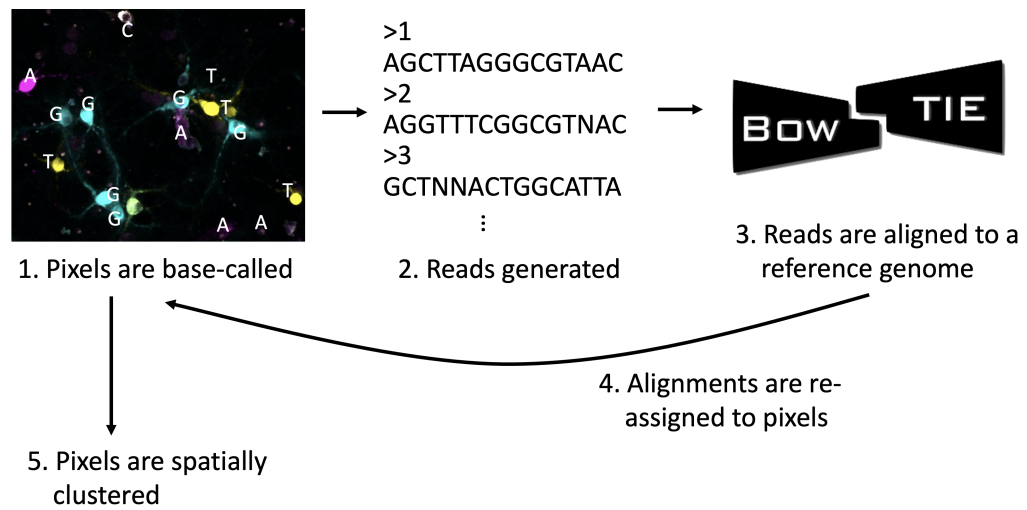


FIGURE 4.9: **Base-calling schematic for FISSEQ.** A schematic representation of the base-calling protocol in Lee et al., 2015.

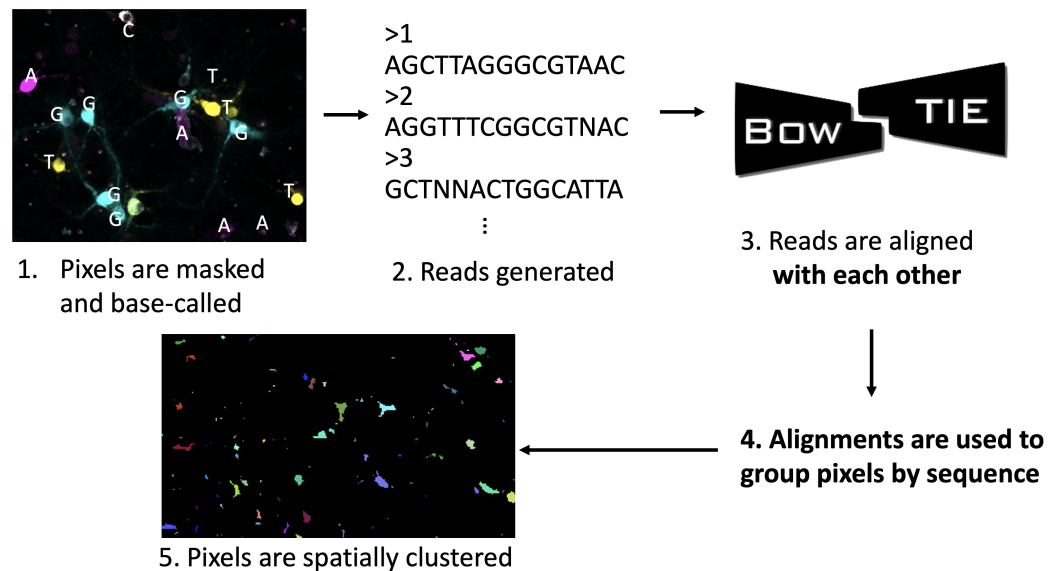


FIGURE 4.10: **Reference-free base-calling schematic for *in situ* sequencing.** A schematic representation of base-calling without a reference genome designed for ISS of random oligonucleotide barcodes.

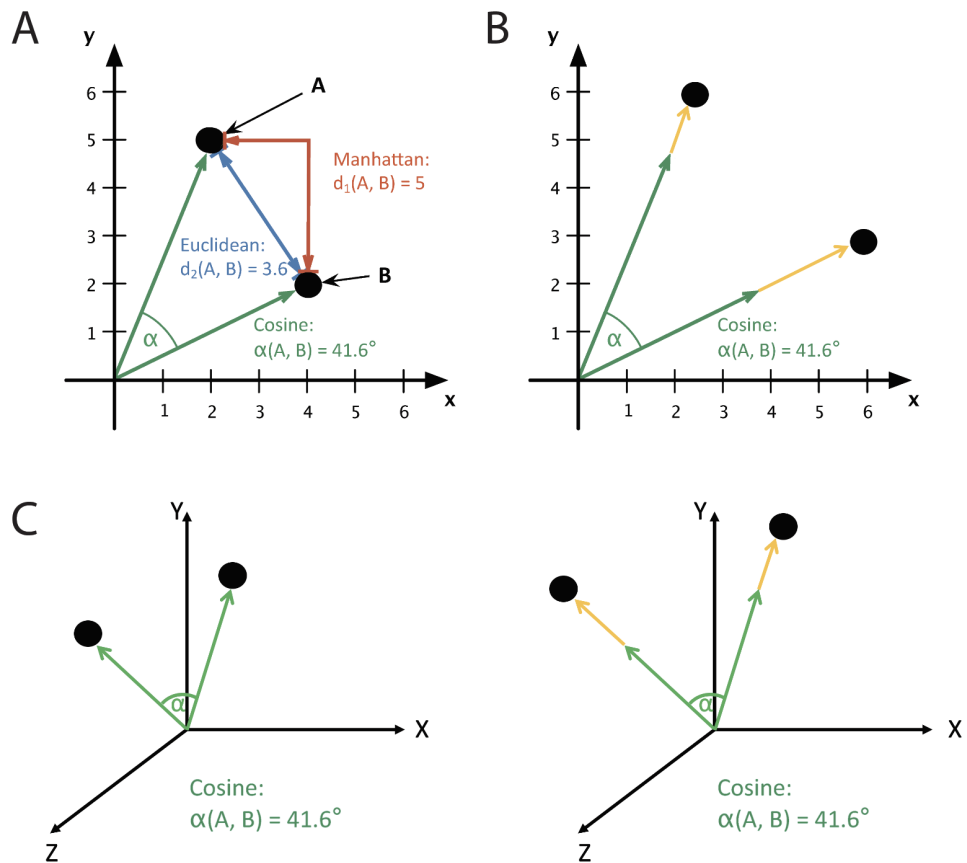


FIGURE 4.11: **Schematic illustrating magnitude-independence of cosine distance.** (A) Schematic of cosine, Manhattan, and Euclidean distance between two points. Adapted from Evert et al., 2016. (B) Schematic illustrating that cosine distance does not change with vector magnitude in 2 dimensions. (C) Schematic illustrating that cosine distance does not change with vector magnitude in 3 dimensions.

library and create a sort of pseudo-reference, we chose to circumvent the need for a reference entirely. The initial sequenced read-outs from the microscopy images are instead aligned to each other and those alignments are used for clustering neighboring pixels with the same barcode (Figure 4.10).

Beyond the removal of the use of a reference genome, we further iterated upon the initial base-calling of the ISS images. Since the images we were working with could have quite strong background noise and fluorescence from the cell nuclei, we determined that a simple intensity threshold would not work. To circumvent this, we decided to take advantage of the fact that real sequencing reads should, except for repeated nucleotides,

change fluorescence colors between subsequent imaging cycles. So, real amplified nucleotides that we are trying to detect, would have a greater distance between cycles than background and nuclei pixels. To capture this distance we used cosine distance which measures the angle between two vectors. Unlike Manhattan and Euclidean distances, for example, cosine distance does not change with magnitude of the vectors (Figure 4.11) (Evert et al., 2016). Calculating the cosine distance for each pixel between imaging cycles we could then threshold for pixels with large cosine distances to identify pixels representing barcodes. We tried a variety of thresholding approaches, and found that adding an pseudo-background to all pixels prior to calculating cosine distance gave the best identification of sequenced reads (Figure 4.12). Trial-and-error selection of a threshold for filtering, found that a lowered threshold further increased signal (Figure 4.13). Finally, after identifying pixels representing ISS reads, each pixel was assigned a base based on the imaging channel with the highest intensity for that cycle. Ties were simply ignored as they generally represented the background with low intensity across all channels (Figure 4.14). Note that each channel is normalized by varying exposure times by the imaging hardware itself.

After initial base-calling, reads generated from the pixels are then aligned to each other using Bowtie (Figure 4.10) (Langmead et al., 2009). Then pixels are spatially clustered using a greedy approach collapsing largest groups first (Figure 4.15). This greedy approach guarantees a Hamming distance of 2 or less within each clustered group. This clustered image then undergoes a morphological erosion using a 3x3 structure (Figure 4.16). Morphological erosion left 223 clustered groups compared to an original 8362 groups. Finally, applying this workflow to a second dataset, we found that this pipeline is widely applicable to datasets not involved in its creation (Figure 4.17). Though some tinkering in thresholds improves the performance.

The base-calling image analysis pipeline for ISS detailed in this section represents one such approach. As previously discussed, there are a variety of approaches across canonical image analysis and bioinformatics that can be used and combined to process ISS data (see Section 2.3). Here, we simply modified an existing workflow to adapt it to ISS of barcodes rather than endogenous mRNA. We further iterated on various steps of

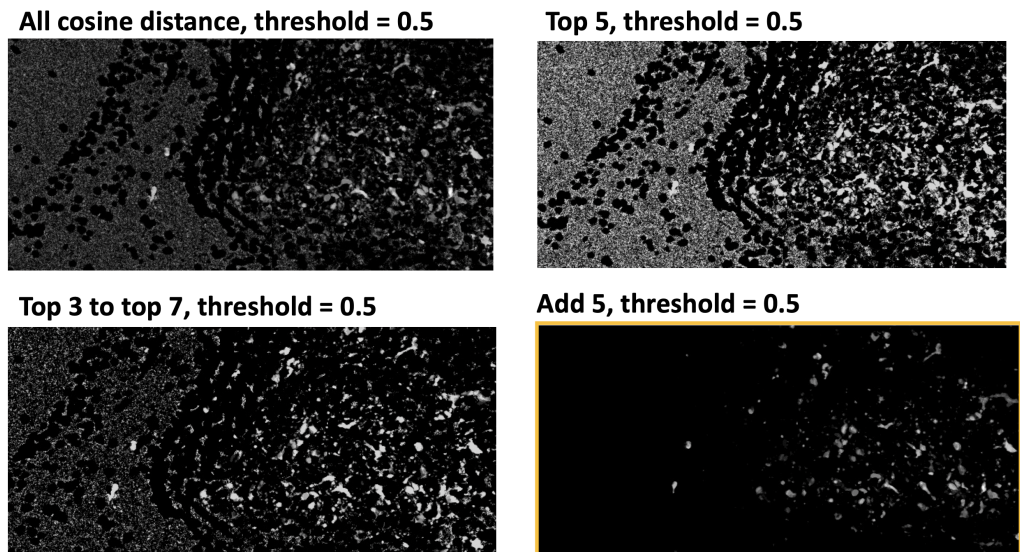
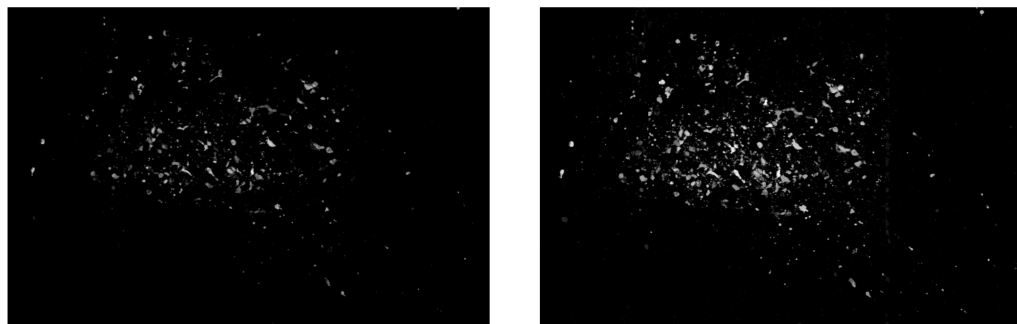


FIGURE 4.12: **Trial-and-error testing of various thresholding approaches to identify pixels corresponding to sequenced reads.** Average of all transitions greater than 0.5 (top left); average of the top 5 cosine distances between transitions greater than 0.5 (top right); average of the top 3 to 7 cosine distance between transitions greater than 0.5 (bottom left); and adding 'pseudo-background' to all pixels before calculating distance, averaging pixels, and thresholding as in top-left (bottom-right). These images show counts of number of transitions above a the specified threshold normalized to grayscale. The best approach, adding a 'pseudo-background' is outlined in golden yellow (bottom right).



Avg. all add 5, threshold = 0.5

Avg. all add 5, threshold = 0.3

FIGURE 4.13: **Trail-and-error determination of threshold for base-calling.** The initial threshold chosen in Figure 4.12 was further adjusted from 0.5 to 0.3 to increase the number of pixels potentially representing ISS reads.

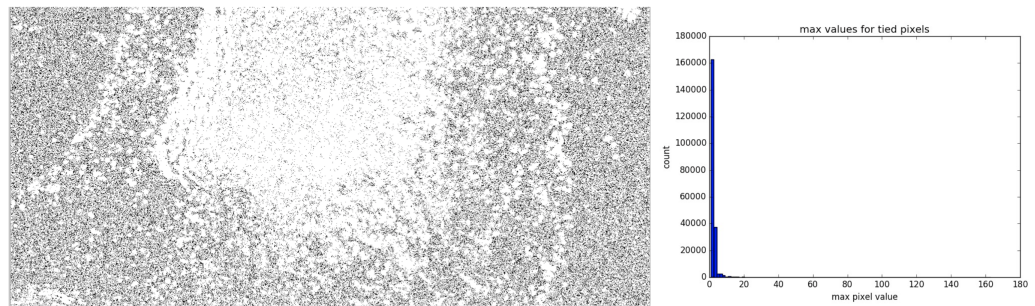


FIGURE 4.14: **Handling ties in maximum intensity for base-calling.** A pixel is assigned a base based on the imaging channel that has the highest intensity. Left: Ties in maximum intensity between two or more channels are shown in black. Right: Histogram of maximum pixel value for tied pixels.

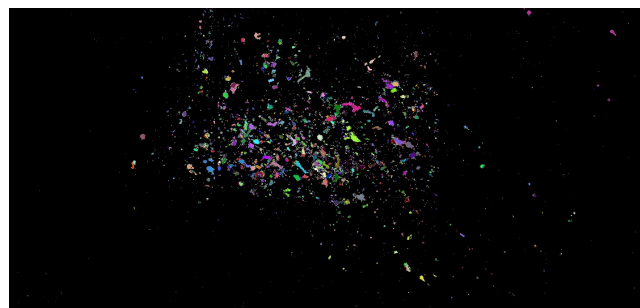


FIGURE 4.15: **Pseudo-colored grouping of pixels sharing the same sequence.** Neighboring pixels with the same or similar (Hamming Distance ≤ 2) sequence are clustered together using a greedy approach that collapses largest groups first.

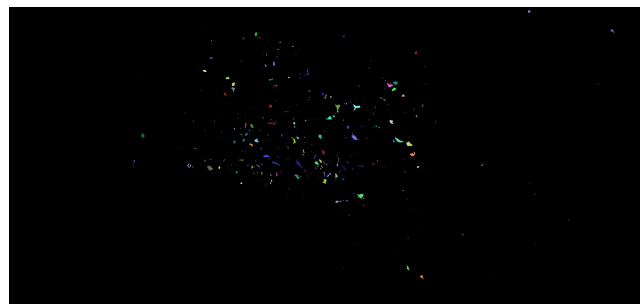


FIGURE 4.16: **Morphological erosion of clustered pixels.** The image shown in Figure 4.15 undergoes morphological erosion with a 3x3 structure to better define cell bodies. Resulting clusters are pseudo-colored (not the same colors as previously in Figure 4.15).

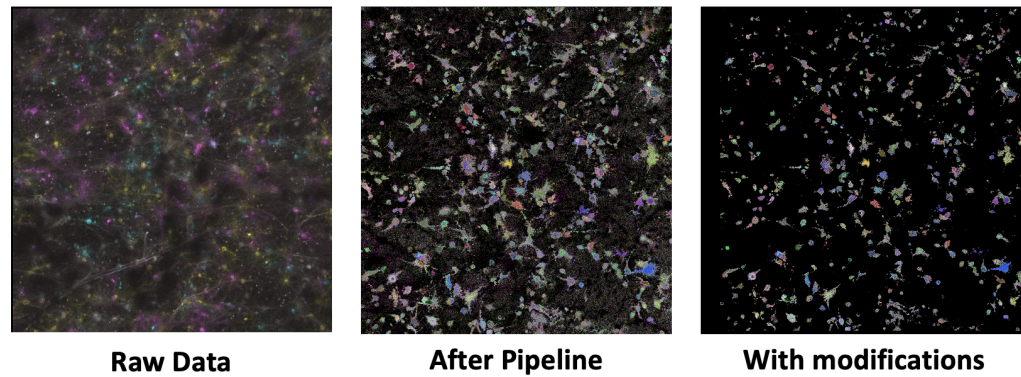


FIGURE 4.17: **Application of base-calling pipeline to an independent dataset.** The base-calling workflow described in this section is applied to a dataset not involved in its creation. Left: raw data; Middle: after base-calling with prior parameters; Right: adjusting parameters to this dataset.

this workflow to improve base-calling for our experimental images. Future work on this pipeline could improve on the masking/thresholding and explore alternate group clustering methods. As ISS methods become more widely adapted, image analysis workflows for base-calling will also become more robust and standardized.

4.5 Discussion

Through three vignettes of using spatially-resolved transcriptomics in neuroscience, we highlight a variety of applications of spatial expression. In contrast to the previous chapter (Chapter 3), which takes a global approach to genes and brain areas, the vignettes described in Section 4.2 and 4.3 focus in on particular genes and brain areas, respectively. By focusing only on the cadherin family of genes in the auditory cortex in Section 4.2, we were able to inspect replicability on the level of comparing ISH and ISS images when quantification of such images did not make sense. Further this focus on a handful of genes, allowed us to re-quantify the ABA images, which is a gargantuan task for all genes and areas. In the next section (Section 4.3), we explored the potential of linking spatial and single-cell data in the motor cortex. While our results show no significant enrichment between cell-type and spatial markers, we suggest that more robust analysis either separating out layers of the cortex or in other brain areas be done before determining whether spatial patterning of expression is being driven by cell-type compositional effects. Finally,

in Section 4.4, we step back and tinker with the initial data processing required for ISS approaches. Adapting ISS for use with MAPseq (see Section 2.4.3) meant that sequenced reads were random barcodes instead of endogenous mRNA. To base-call these barcodes, we modified and iterated on an existing workflow to process ISS images. That these vignettes span data pre-processing, replicability, and cross-technology integration illustrates that there are many exciting research questions and types of analysis to pursue withing the umbrella of spatially-resolved transcriptomics and neuroscience.

Chapter 5

Conclusions and perspectives

5.1 The potential of spatially-resolved transcriptomics

Spatially-resolved transcriptomics is a recent, exciting development in biology. In the last 5 years, new technologies were constantly being developed and improved upon (see Chapter 2). Today, spatially-resolved transcriptomics has just begun to mature and is increasingly becoming readily available to most laboratories. The recent wave of single-cell technologies lacked information on the spatial origin of the cells they are sequencing. Spatially resolved methods are perfectly poised to fill this gap and allow us to answer questions about the spatial patterning of cell-type specific expression. This is powerful. Further, in the brain, spatially-resolved expression can link molecular (e.g. epigenetics, expression, etc.) and mesoscale properties (e.g. projections, morphology, etc.) of the brain, enabling a multi-modal approach to neuroscience. Beyond neuroscience, the availability of spatially-resolved gene expression assays enables us to ask and answer open questions in all fields of biology. For example, spatially-resolved transcriptomics will allow us to capture at a large-scale changes in spatial patterning of expression in development (Di Bella et al., 2021). Spatially-resolved expression is ripe to usher in the next wave of discovery (Marx, 2021).

5.2 Summary and conclusions

With the recent development, explosion, and high potential of spatially-resolved transcriptomics tools, understanding the robustness of these tools is crucial. In this thesis we sought principally to understand how robust these new technologies are relative to each other and more traditional spatial approaches (e.g. *in situ* hybridization). Taking advantage of the adult mouse brain as a highly-studied tissue with stereotyped sub-structure, we examined the replicability of a new spatially-resolved transcriptomics tool (ST; see Section 2.2.1) to more traditional ISH collected by the Allen Institute (see Section 2.4.1). While replicability of individual techniques are always compared to previously published methods, this usually occurs only for a handful of genes or spatial areas. To our knowledge, our work in Chapter 3 is the first whole-transcriptome, whole-brain assessment of the replicability of spatially-resolved methods across independent datasets and platforms. In general, we found that biological conclusions drawn from the two datasets were replicable. In other words, we could distinguish between spatial samples based on their transcriptional profiles. However, when we tested models trained in one dataset in the opposite dataset, the results did not always hold up. Specifically models trained in the ST dataset would generalize to the ABA, but the reverse was not true.

We suspected that some of the lack of generalizability in Chapter 3 was not due to the underlying data itself or the biological differences between individual mice, but rather the pre-processing of the ABA data specifically. In Chapter 4 we sought to re-quantify a handful of the ABA genes in only one area (the primary auditory cortex) to compare to *in situ* sequencing (BARseq2) data (see Section 4.2). Here, we chose to re-quantify the raw ABA data from images instead of working with the provided values as above to circumvent the potential lack of generalizability arising from pre-processing of the ABA data. We found that for these genes, the re-quantified ABA data does indeed replicate with BARseq2 ISS.

Exploring other applications of spatially-resolved transcriptomics in neuroscience, we took a first stab at linking the ST and ABA data (as in Chapter 3) with single-cell data

(also in Chapter 4). We asked if spatial patterns of expression are driven by differential cell type composition in the primary motor cortex (see Section 4.2). To address this we compared pre-defined cell-type markers with spatial markers of the MOp, but did not find a significant enrichment. We caution that this is only preliminary work in one brain area and suggest that using a cortical area as a whole, where laminar expression is more distinct than cortical area expression, could not be the ideal first test case to answer our question. Finally concluding Chapter 4, we share the development of an image analysis pipeline to base-call ISS reads adapted for sequencing of a random barcode library used to trace neuronal projections (see Section 4.4). This pipeline is able to base-call sequencing images, including images not used to develop the workflow, though manual tuning of parameters is required to optimize it.

5.3 Future directions and broader impact

We hope that our work in this thesis lays a foundation for the continual assessment of replicability within spatial data and across modalities. The obvious, and perhaps most critical, extension of this work is the continual integration of whole-brain, whole-, or nearly whole, transcriptome spatially-resolved datasets in replicability studies. Additional datasets would help clarify some of the ambiguity in interpreting the results in Chapter 3. Since there is a discrepancy in cross-dataset generalizability, the addition of further datasets could help determine which of the two most closely represent actual biology. Additional, more specific extensions of this work such as re-quantification of the central ABA resource or the production of a spatial gene marker database was discussed previously in Section 3.5.

Building on some of the work in Chapter 4, we propose that the full potential of spatially-resolved transcriptomics is in tandem with multi-modal integration. As introduced in Chapter 2, there are now a wide variety of computational tools that allow for integration of spatial data with other types of molecular data, such as scRNA-seq or single-cell chromatin profiling. In neuroscience, spatially-resolved expression has been integrated with other modalities beyond molecular data. For example, spatial expression and neuronal

projections have been studied together (see Section 2.4). Today, there are many exciting emerging experimental and computational tools assaying a variety of biological phenomena that will allow us to have a comprehensive, multi-modality understanding of biology.

Beyond the scope of this thesis, robust spatially-resolved transcriptomics tools have the potential to make an impact in a variety of biological fields. For example, spatial data can help biologists and clinicians understand tumor heterogeneity. We can now assay all, or many, genes in such a manner and create expression-based histology-like images. Spatial tools can contribute to the study of development, where spatial patterning of developmental genes is known to be critical to a variety of animal body plans. Outside of multicellular organisms, spatial expression tools can even help us understand dynamics in colonies of unicellular organisms such as bacterial plaques. Between organisms, spatially-resolved transcriptomics can potentially answer questions about inter-species dynamics such as in the study of gut microbiota. The applications and potential of spatial techniques across biological questions is broad.

5.4 Biology and big data are one in the same

We prefaced this thesis with a discussion on the role of consortia and big science in biological research today (Chapter 1). As we consider the future of spatially-resolved transcriptomics and its interconnectedness with other data modalities, it is clear that the volume and breadth of research enabled by these tools will also involve big biology approaches. In fact, some current spatially-resolved tools were developed as parts of larger consortia-based research (see Section 1.6) and wide adaptation of these tools is dependent on commercialization or use by high-profile and/or large numbers of labs (i.e. consortia). Today, as both the scope of individual datasets and the quantity of datasets increase, big data and big biology approaches are becoming synonymous with modern biology research. Machine learning and other computational tools, that historically relied on large sample sizes, are becoming increasingly useful in biology. One relevant example is the adaptation of image analysis and computer vision for spatially-resolved transcriptomics image analysis. Though certainly not exclusively, the future of spatially-resolved transcriptomics, and

biological research more broadly, is intertwined with big data and big science approaches as we seek to understand increasingly complex open questions on dynamic interactions and emergent properties of biology.

Appendix A

Unpublished Review: Consortia

This appendix contains the current full text of an under preparation review/perspective piece, which was authored jointly by Shaina Lu, Nathan Fox, Anthony Zador, and Jesse Gillis. I wrote the original text presented here.

Introduction

Twenty years ago, at the start of the millennium, former United States President Bill Clinton and former United Kingdom Prime Minister Tony Blair jointly announced the completion of the first sequencing of the human genome (cite). It would be a few more months until the draft genome was published and a couple more years before the completion of the final version. While few scientific enterprises conclude with the fanfare of the Human Genome Project (HGP), this moment marked a milestone for all biological and biomedical research. Not only would the research findings of this consortium become foundational for modern research in these fields, the technology developed, standards- formal and informal- set, and style of research would transform modern biology. This was the start of consortia-based big science in biology.

Today, in the nearly two decades since the HGP announcement, consortia of all types permeate all corners of biology. These big science approaches have generated previously unthinkable datasets: the Allen Institute's brain-wide spatial expression atlases, the extensive characterization of DNA by ENCODE, and the organized functional gene annotations of GO- to name a few. They have set forth laudable data sharing and ethics principles: the public data sharing guidelines of the HGP's Bermuda Principles and the establishment of the Sequence Read Archive (SRA) in support of the 1000 Genomes Project and Human Microbiome Project. They have democratized tools for research: sequencing technologies from the HGP and STAR, the sequence alignment algorithm from ENCODE. Beyond these tangible contributions to science, consortia also provide inspirational value to humanity by increasing public trust in science and promoting cross-cultural, international collaborations.

Despite all of these contributions to biology, the consortia-based approach is not the be-all and end-all of research. As with any large organizations, consortia can promote group mindsets and hamper creativity. Smaller, dynamic groups may find it hard to compete with resource-rich consortia further propagating the Matthew's effect present in science. Further, consortia often focus on better-established, mainstream research topics, leaving oddball ideas to smaller groups (Bhattacharya and Packalen, 2020 ?). Research on these oddball ideas have proven pivotal in the past. (Famously, the discoveries of Taq polymerase and CRISPR each came from research on the Yellowstone hot springs and on yogurt, respectively). Additionally, trainees involved in large-scale collaborations may find it difficult to get the recognition necessary for scientific career advancement. These are just a few of the possible concerns with big biology.

We stand at the cusp of a potential new era for large-scale collaboration in biology. While fields such as genetics and genomics now have a two decades rich history, other fields like neuroscience are newly diving deep into these big biology approaches. In 2013, Sean Eddy penned an eloquent essay pointing to successes and failures of ENCODE, largely attributing negatives to a failure to properly categorize what type of consortia ENCODE is and what its findings contribute to biology- more on this later (Eddy, 2013 Curr Bio). Eddy ends his essay with a plea to do better for the next big consortia in neuroscience. Today, the BRAIN initiative and the International Brain Laboratory, two large consortia in neuroscience, are in full swing. Further contributing to potential paradigm shifts include (1) new funding initiatives such as the Chan-Zuckerberg Initiative funding the previously overlooked development and upkeep of scientific software and (2) the pre-print server bioRxiv shaking-up publishing. Finally, COVID-19, leaving no stone unturned, will also fundamentally change the way and speed life science research is done. With this new era of collaborative research, we must critically examine the organization, outputs, successes, and failures of past consortia to avoid past downfalls.

Often, scientists like to think of themselves as above the fray, but to understand any group of scientists working together toward a common goal is no different from understanding a group

of non-scientists working together (credit to Lauren Wool). The study of groups of people working together is well described in the social sciences, and particularly, for organizations, in economics. Here, we attempt to borrow from the language of economics to characterize consortia of all types in biology. We begin by attempting to define what constitutes a consortia, move to exploring the products of these collaborations and how to quantify them, discuss the monetary system of consortia, consider whether consortia and their products are public or private, and finally explore harmful and beneficial aspects of consortia. As scientists continue to organize into different types of groups and these groups proliferate, we can now take pause to assess these different structures and their efficacy using the framework of economics.

What is a consortium?

Diversity of consortia

Even before modern-day consortia, humans have long been organizing themselves behind common scientific goals. During the age of exploration and beyond, determining a ship's location on long ocean voyages was key. While latitude was easily found by tracking the sun, longitude was a notoriously hard problem. To incentivize this, prizes were offered by European rulers as early as the mid-16th century and as recently as the longitude rewards of the British government established in the early 18th century. Also in the 18th century, when Italy was still a bunch of fragmented states, scientists across what would later become a unified Italy, decided they wanted to find eel gonads and solve the mystery of their reproduction (radiolab). The question of eel reproduction was so perplexing and long-standing, that they believed this discovery would be a part of an unifying Italian national brand. A more recent example of big-science is the space race of the US and Russia in the cold war era. In the US, the multi-state and multi-billion National Aeronautics and Space Administration (NASA) was created. Incentivization, collaboration, and competition through large-scale scientific efforts has been a part of human civilization for centuries.

In biology today, consortia come in all shapes and sizes. These consortia vary not only in their research focus, but also in how they are created, organized, and funded- to name a few axes of variation. Take for example, the Critical Assessment of Structure Prediction (CASP). CASP is a long-running (biannually since 1994), grassroots collection of scientific groups working toward the common goal of predicting unknown protein structures with computational modeling. At its core, it is a contest where the individual groups work in competition with each other to build the best structure predictors. (This competition has been in the news lately because of the complete domination of its central task by DeepMind's AlphaFold2 (cite Mohammed AlQuraishi blog, etc.).) In contrast to CASP, are highly-centralized consortia, such as the BRAIN Initiative Cell Census Network (BICCN) arm of the broader BRAIN initiative. The BRAIN initiative came about as a legacy project launched and funded through a US Presidential initiative of the Obama administration. The BICCN faction has the central goal of creating comprehensive cell type atlases across model organisms and humans. This consortium is a goal- and funding- driven initiative very dissimilar to CASP.

A third class, tangential to both CASP- and BICCN-like consortia, are consortia organized around longitudinal data collection. A couple of famous examples of this are the UK Biobank and Framingham Heart Study. These types of consortia run over long periods of time to collect longitudinal data that would likely not be possible without the respective consortia's existence. The Framingham Heart Study has been running since 1948 in its namesake town of Massachusetts (Andersson *et al.*, 2019, Nature Reviews Cardiology). This study now includes over 14,000 participants spanning three generations and is responsible for much of our modern-

day understanding of the risks and prevention of cardiovascular disease (Mahmood *et al.*, 2013, *The Lancet*). Not unlike BICCN, the Framingham Heart Study was started through the National Heart Act signed by former U.S. President Harry Truman in 1948 (Mahmood *et al.*, 2013). (Truman was vice president to Franklin D. Roosevelt who suffered from largely undiagnosed cardiovascular disease.) While younger in age, the UK Biobank is also a longitudinal study. Started in 2006, the roughly 500,000 participants agreed to be followed for at least 30 years (Bycroft *et al.*, 2018; *Nature*). The scope of this data is extensive ranging from simple survey demographics, MRIs of the brain and heart, and genotyping of blood samples. While some of this data collection is still on-going, this dataset has already proved invaluable (some representative/editorial examples). Presently, the UK Biobank resource has also enabled researchers to name putative risk factors of COVID-19 (Armstrong *et al.*, 2020 microbial genomics; Yates *et al.*, 2020 primary care diabetes;) and to identify race and socioeconomic demographics of COVID-19 infections (Niedziedz *et al.*, 2020; *BMC Medicine*). These large-scale data-based consortia enable well-supported, population-level studies both in the world that conceived the studies and well beyond. They enable rapid research and understanding in real-time crises (a la COVID-19).

Conceptualizing consortia as economics agents

Looking at these historical and modern examples, it is easy to see that the label 'consortia' can be a bit of a catch-all term encompassing much diversity. With this broad variation, can research consortia be neatly defined and categorized to better understand them and their outputs? When faced with a similar problem in defining what constitutes a company (usually for taxation), many governments have come up with elaborate frameworks. Taking the United States as an example, and not even considering state-level frameworks which vary between states, the federal government can organize a company into one of 4 categories. Each of these in-turn has additional qualifying frameworks and labels that can be used to further categorize companies. In actuality, this taxonomy is further complicated since companies, with a few exceptions, are incorporated on the state level. The result is a messy, multi-layered categorization full of special cases and exceptions. For both companies and consortia, distinguishing and labeling is messy; often our labels are intertwined, overlapping, and incomplete. Broad categories, such as those proposed through exemplars above, can be easier to understand, but looking closely, these taxonomies are often unresolved.

Despite the diversity of consortia obfuscating a clear taxonomy, it is clear that the various consortia simply represent classes of coordination. A well-trodden path in thinking about coordination and its variations is through the economics framework of firms. The theory of the firm encompasses many ideas in economics that seek to explain why firms would exist and how they would work in contrast to the open market which was the predominant focus of economic theory at the time. Ronald Coase's "The Nature of the Firm" is recognized as among one of the first conceptualizations of firms (Coase, 1937). Coase posited that firms exist simply because organizing and negotiating a contract for every need on the open market becomes too costly. We can draw a parallel to scientific consortia where we consider a single consortium as a firm and labs under one principal investigator as an individual on the free market. Individual labs often collaborate without any formal arrangement, though there are exceptions to this such as working with human data or in some international collaborations (Risa). Further, long-term collaborations can often lead to joint grant applications that in some cases can represent funding opportunities from or participation in a consortium. In this analogy, we posit that consortia exist to formalize interlab collaborations without the need to identify successful collaborators, process cumbersome 'contracts' each time, and ease barriers to the work itself (i.e. allowing unbarred data sharing).

Conceptualizing consortia as an economic firm can help us to understand how they work and their contributions.

What are the products of a consortium and how do we measure them?

Variety in consortia outputs

In keeping with the analogy of the firm, a consortia is organized to produce scientific goods, but what exactly are those goods? As previously introduced, consortia have made a variety of contributions to the larger research landscape including, but are certainly not limited to: data, data-sharing infrastructure, research standards, democratizing new technologies, computational software, and of course, scientific knowledge communicated through papers. In this section, we will explore each of these outputs through examples.

Data. In modern biology, with the competition/benchmarking based groups as an exception, one of the more ubiquitous outputs of research consortia is data. The flagship output of the HGP is the reference human genome which expanded into cataloguing all functional elements of the genome in ENCODE and model organisms in modENCODE. As sequencing prices dropped, many consortia sprung up around cataloguing the genomic diversity of, often previously underrepresented, human populations: the 1000 Genomes Project (now, The International Genome Sample Resource) sequencing individuals across the world in an effort to identify rare variants, UK10K in sequencing 10,000 UK individuals to identify rare variants, GenomeAsia 100K Project in wanting to add diversity to genome datasets by sequencing across Asia (GenomeAsia100K Consortium, 2019), and perhaps most recently, the All of US precision medicine group from the National Institutes of Health (NIH) seeking to gather biological and health data from over 1 million U.S. participants. Similarly, data-focused initiatives also grew around specific interest areas of biology: The Cancer Genome Atlas Project (TCGA) sought multi-omic profiling of 20,000 cancer samples with matched healthy samples, the Allen Institute (itself more of a research institution than consortia) created a comprehensive *in situ hybridization* based spatially-resolved transcriptomic atlas of the developing and adult mouse brains (among many other atlas style resources), and the Genotype-Tissue Expression (GTEx) consortia matching genotype with tissue specific transcriptomics. As technologies were developed and improved, additional data-generating groups utilized these new tools: single cell mouse brain atlases of the BICCN and Allen Institute and the tissue specific single cell atlases of the larger Human Cell Atlas (HCA) and Chan Zuckerberg Initiative's (CZI) Tabula Muris. Finally, it is worth re-highlighting as previously discussed that longitudinal datasets requiring multi-generational organization are mostly impossible outside of consortia. (Richard Lenski's directed evolution experiment in bacteria is a major exception.) These examples are by no means all encompassing, but do illustrate the broad focus on data for many consortia.

Concerning human data, there is a notable lack of diversity in these datasets. For example, not only are the subjects of the Framingham Heart Study predominantly white and of European origin, but this demographic make-up was further claimed to be representative of the 1940s U.S. when the study started (Mahmood 2013), which is simply not true. GTEx, as a recent example, is 84.6% white and 67.1% male (<https://gtexportal.org/home/tissueSummaryPage>) which is not representative of the racial and gender make-up of its host country, the US. Lack of diversity in datasets is not only harmful to the communities that these consortia fail to serve, but also harmful to the research itself. Concretely, diverse human datasets would for example allow researchers to identify more polymorphisms and more generally allow for more robust and generalizable research findings. Relatedly, there has been a recent resurgence in improving the human reference genome to be more representative of the human population and not just consisting of

mostly one single individual as the dominant reference is today. One solution is a consensus genome that would be able to harness the diversity in sequenced genomes to build a better reference (Ballouz *et al.*, 2019 Genome Bio). Recent consortia promise to increase the diversity of human datasets such as the GenomeAsia 100K project, All of Us, and the HCA. As the vanguard for large-scale data generation, consortia are continually responsible for ensuring the diversity of their human datasets.

Data sharing. In many cases, some of these large-scale and/or long-scale datasets likely only exist because of consortia-like effort. However, relative to other highly collaborative fields, high-energy physics for instance, biological data is extremely fragmented (Bonnie Berger at ISMB 2019; Helio/John Hover). With the diversity of data types and constantly evolving technologies, data and associated meta-data in biology is extremely messy, often with inconsistent formatting. In efforts to combat this, there are laudable data-sharing policies and infrastructures that grow out of consortia. Starting again with the HGP, this consortium led in open science by requiring sequence data to be rapidly released prior to publication as laid out in the Bermuda Principles. This set the standard for genomics research, which continues to be one of the most open sub-fields of biology; the genomics field was one of the earliest adopters of bioRxiv pre-printing, having some of the highest numbers of pre-prints (the bioRxiv talk from the CSHL press at in-house). In the early 2000s, out of community demand, the National Center for Biotechnology Information (NCBI), a branch of the NIH, established the Gene Expression Omnibus (GEO) for the main purpose of sharing gene expression data in the form of microarrays (Edgar *et al.*, 2002 NAR). Impressively, GEO continues to be a major resource for sharing multi-omic data today. As previously mentioned, later in the 2000s, the NCBI, in collaboration with the European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ), established the Sequence Read Archive (SRA) in support of international consortia: the Human Microbiome Project and 1000 Genomes project (NAR 2012). The SRA is still used today as one of the main repositories for sequencing data. Recently, as a part of the BRAIN Initiative, the Neuroscience Multi-omic Data Archive (NeMO) was introduced as a multi-modal data repository for modern neuroscience datasets spanning physiology, genomics, and beyond.

These days, however, as datasets continue to proliferate in both quantity and individual size, data availability is often not enough to render them useful to the research community. One barrier to accessibility is poor metadata, a problem recently compounded by single cell sequencing platforms which GEO was not designed to support. Recognizing this, the BICCN, for example, is currently reckoning with how to make their data continually accessible and easy to use. A part of the BICCN collaboration, the Karchenko Lab has proposed a Cell Type Annotation Platform (CAP) to standardize the sharing of single cell data and associated metadata. Also going further than simply serving data, there have been recent efforts to pre-process large bodies of data using the same bioinformatic pipelines such as the Genome Data Commons (GDC) of the National Cancer Institute (NCI), which has consistently processed data across many large cancer genome datasets including the aforementioned TCGA.

Beyond the influence of consortia efforts, there are additional concerns and barriers to data sharing. With human data, ensuring the safety and privacy of donors is paramount. Sometimes, however, data protection regulations can inadvertently inhibit scientific data sharing. One example is the European Union's General Data Protection Regulation (GDPR) which was passed in 2016 and went into effect in 2018. While the target of the regulation was to protect personal data, genomics data on human subjects can sometimes fall under this regulation making it challenging for genomic and health data sharing both within and beyond the EU (PHG report, Robert Eiss in Nature, Molnár-Gábor and Korbel, 2020 EMBO Mol Med). In the two years since going into effect, frustration in the GDPR's lack of clarity in interpretation abounds and

international collaborations, including consortia, have been stalled (Robert Eiss in Nature, Molnár-Gábor and Korbel, 2020 EMBO Mol Med). Beyond the EU, researchers who want to work collaboratively on human data across international borders may need official appointments or contracts drawn to even access the data (Risa). While these protections on human data are ultimately good, setting up new collaborations can become costly much like the cost of setting up new partnerships on the open market. Having established consortia, like firms, can help alleviate these barriers to working collaboratively on human data.

Research standards. Having power in numbers, and often funding, consortia have the leverage to set research standards for the field. An obvious example of a research standard is the human reference genome first published by the HGP. The importance of having an accepted standard reference in genomics research can be stated by analogy to the need to have a reference for a standard measurement, much like the recently redefined official kilogram reference (Stock *et al.*, 2017 Metrologia). Another example of a standard reference is the whole-transcriptome *in situ* hybridization atlas and common coordinate framework from the Allen Institute. The former is heavily used as a comparison for nearly all subsequent spatially-resolved transcriptomics datasets in the mouse brain and the latter provides a standard coordinate system for areas of the mouse brain (cite).

Beyond providing a reference, the entire purpose of some consortia can be to define research standards. For example the MicroArray or subsequent Sequencing Quality Control (MAQC/SEQC) effort from the U.S. Food and Drug Administration (FDA) is entirely organized around benchmarking transcriptomics technologies so that they could be reliably used in diagnostic and regulatory applications (citation). A second, slightly different, example of a research standard is the controlled vocabulary to define functional genes created by the Gene Ontology (GO) (citation). With the tagline goal of “Unifying Biology,” GO has enabled standardized annotations of gene function and the easy assignment of functional enrichment for any given gene. Whether an offshoot of data generation or the central goal of a consortium, consortia often play a large role in defining and setting research standards.

Democratizing new technologies and computational software. Being at the forefront of data generation means that consortia are often also leaders in the creation and democratization of technological development. In order to achieve the goals of a consortium, new technology needs to be pushed to production scale in contrast to the proof-of-principle style of individual labs. Again, the HGP provides a touch stone example, in that sequencing of the human genome started out using the laborious bacterial artificial chromosome sequencing (BAC-sequencing) technique where fragmented human DNA was cloned into bacteria for replication then sequencing and re-assembly. By the end, through Craig Venter’s group that split off, the human genome was competitively sequenced using shotgun sequencing, a more efficient sequencing approach that eliminated the bacterial cloning step and laid the foundation for future sequencing (cite). The development and commercialization of production-scale sequencers can be viewed as a direct outgrowth of the HGP. A more recent example is that of the International Brain Laboratory that sought to standardize neuroscience experimental setups and behavioral tasks which are generally bespoke to individual labs (IBL *et al.*, 2020 bioRxiv). In the process, hardware components to assay rodent behaviour were further developed and made available (Sanworks).

These technological advancements are not limited to hardware; many popular scientific software have grown out of consortia as well. For example, the popular RNA sequencing read alignment algorithm Spliced Transcripts Alignment to a Reference (STAR) was developed to align reads generated from ENCODE (Dobin *et al.*, 2013 bioinformatics). STAR is now widely taught and used in bioinformatics courses and research, respectively. In keeping with the advancement of sequencing platforms, subsequent versions of STAR for single-cell RNA sequencing have been

developed (STARsolo preprint). A second example is the SpaceTx pilot project of the HCA which sought to benchmark and streamline the analysis of spatially-resolved transcriptomics. This resulted in the Starfish suite of tools to analyze spatial expression datasets. The SpaceTx effort even included a hackathon to bring together the community in working on this tool (nature feature).

In many ways, outside of commercialization, the maturation of technologies in the academic realm is only possible through consortia efforts. Outside of consortia, funding to develop technology beyond a prototype is scarce in current science funding models (Tony). Consortia can provide the research dollars needed to mature a technology. With so many bespoke technologies in individual labs, one way to think of the relationship between consortia and technology development is in reference to Paul Krugman's theories on international trade and economies of scale (cite). Grossly simplifying, Krugman proposed that based on economies of scale, it is cheaper to mass produce a product. Generally, mass production would reduce individual prices, but global trade mediates this. We can abstract this further as a hub and spoke model, where in consortia we have a hub, either a single lab or whole consortia, that has the expertise to run a complex technique and provide it as a service to other research groups, the spokes. Another phenomenon related to consortia and technology worth noting, is that sometimes the adoption of a technology by a consortium can cause a competing approach to be sidelined for little apparent reason.

How do we measure the impact of a consortium? (Consider only papers, here)

In a standard company, products can be straightforwardly described through quarterly sales and earnings reports. The outputs of consortia, however, are usually not directly sold or monetized. Combined with the variety of consortia products, the broader categories of which themselves are not comprehensively enumerated above, it can be difficult to quantify the impact of various consortia. In this section we will consider papers, given that scientific knowledge communicated through papers is the ultimate consortium product, as a metric for quantifying consortium output. (Since every consortium ultimately results in publication(s), we will not enumerate examples of consortia producing papers.) Further, other consortium products such as technology, software, and data are often described in publications. Though, there are important caveats to this to note; for example, updates to software that may require many research hours often do not get a new, independent publication (citation).

The use of publications as a measure of research output is not a new idea. A variety of bibliometrics such as the h-index have long sought to quantify research productivity through papers. The h-index represents the number of articles an author, or research group, publishes that has been cited at least that same number of times (Hirsch 2005, PNAS). There are variations on this such as a time limited h-index that only includes publications from a predetermined recent number of years or Google Scholar's i10-index that simply reports the number of papers an author has published with at least 10 citations. However, the h-index and related bibliometrics are deceptively simple. These values, while useful, have serious limitations (citation). They are calculated based on publication databases such as Google Scholar and Web of Science, which means they can be biased based on what publications are included in the databases. Indices can also include conference abstracts in their calculations, which is important for some fields like machine learning, but might artificially inflate h-index in others like most biology sub-fields. Indices also vary widely across fields, making generalization hard. Knowledge of publication practices of specific fields and sub-fields can help mediate these pitfalls, but as research and resulting publications continue to become increasingly interdisciplinary this will only help so much.

In recent years, there have been proposals for new bibliometrics that seek to iterate on the h-index, in some cases by including other metrics of engagement and interaction with a publication beyond citations. While still rooted in citations, variations of an index based on Google's PageRank algorithm, propose using the algorithm to determine the impact of publication in a citation network, instead of just the citation count (Senanayake *et al.*, 2015 PLOS One; Gao *et al.* 2016; PLOS One). Other proposals have advocated for including broader engagement with papers and preprints such as social media interactions (Carlson and Harris, 2020 Plos Bio; Díaz-Faes *et al.*, 2019 PLOS One). Many of these non-citation bibliometrics are tracked through Altmetric and their Attention Score (citation), but the h-index and its variants remain dominant indicators of research impact.

Further compounding potential drawbacks of bibliometrics, consortia can have unique publication and authorship practices relative to each other and individual labs of the same field. In biology, for the most part, author order carries meaning. First author(s) are responsible for the bulk of the work presented in the paper and putting together the manuscript. Last author(s) usually represent principal investigators (professors) who secure the funding, provide mentorship and guidance on the project, and ultimately run the laboratories the research was performed with. Middle authors contribute to the research presented in decreasing order of importance from the first. Sometimes in biology, for better or worse, even the order of co-first authors carries significance. Consortia authorship on the other hand, can list the consortia as first, last, or only author; have alphabetical authorship, or even in some cases, random author order. This heterogeneity can make it difficult for those outside the consortium to straightforwardly assign credit, which is key to the success of scientists for future positions and funding, to those involved in consortia and existing bibliometrics fail to capture these nuances.

On the topic of credit assignment to individual researchers involved in consortia, there are many examples of Matthew's effect, often expressed as "the rich get richer and the poor get poorer," at play. Contributing to research consortia can inflate an individual's h-index, even if the contribution is minimal compared to driving a project of their own. Further, consortia themselves usually have bargaining power with publishers that compete to provide a consortium with packages of papers guaranteed to be published in their (usually high impact) journals. These are mutually beneficial relationships where consortia papers tend to be highly cited driving up journal impact factor, a ratio of citations to number of papers published, and consortia essentially get to place their research papers in prestigious journals. As foreshadowed with technologies above, these effects greatly serve those in consortia, while potentially harming individual labs that do not participate. Publications and citations are, for better or for worse, extremely important to individual researchers for getting future jobs and research funding. Finally, with so much emphasis on using papers for measuring impact, consortia can fall into the trap of Goodhart's law, cited as "When a measure becomes a target, it ceases to be a good measure" (Strathern 1997). For example, many consortia famously publish proliferately on the same core subject, such as publishing a Users' Guide on a resource created and sometimes separately published.

There are no easy solutions for better ways to measure productivity in consortia. For now, papers and their citations, seems to be the only remotely uniform way to quantify outputs. However, with all the heterogeneity in publication habits, interpretation of these metrics should always be done with discretion.

Do consortia and their products belong to the public?

In the recent race for a COVID-19 vaccine, many pharmaceutical companies received large sums of public funding to promote research and development of a vaccine, such as through

Operation Warp Speed in the U.S. Despite public money invested in initial development, the intellectual property of the resulting product often belongs to the private companies, with governments having to buy vaccines from the same companies, albeit often at a discount, that they initially funded. This kind of agreement brings to light the tension in public-private partnerships, which often provide a much better return on investment to the private parties than the public.

A similar public-private tension also happens with consortia. Again, starting with the HGP; this was a consortia effort that ultimately cost 2.7 billion dollars of tax-payer money set aside by U.S. congress with additional funding from the Wellcome Trust of the United Kingdom (cite). While the research of the HGP has led to innumerable economic and health benefits to the larger society, it is important to note that at the time that the HGP was formally launched in 1990 it was still unclear whether the main product, the sequenced human genome, would belong to the public or not. As previously mentioned, the sequencing of the first human genome, was famously done by both the public HGP effort and a splintered-off private company, Celera Genomics. Celera had applied to patent gene sequences. Negotiations between the public and private efforts rose to the level of involving former Prime Minister Tony Blair and former President Bill Clinton, who jointly declared that “the human DNA sequence and its variations, should be made freely available to scientists everywhere,” (old white house page, news articles) rendering the sequenced genome unpatentable. Luckily, here, public funding led to a (mostly) public effort with a product ultimately available to that same public.

These examples raise broader questions of whether consortia efforts ultimately benefit the public. Are consortia a good use of public funds? Do the products belong to the public? What happens when research done with public funding is translated to private commercialization? The direct benefit of consortia to the public may not always be straightforward and answers to these questions are not easily found, but according to the endogenous growth theory of macroeconomics, investment in technology is thought to lead to overall economic growth (Paul Romer 2018 Nobel). Whether the public sees a benefit of that growth is a wholly separate issue far beyond the scope of this piece. Tangentially, mostly lacking within consortia is the use of contracts. Contracts may be hard to define as active discovery and innovation is happening, but the use of contracts can help navigate conflicts of interest within and beyond a consortia to avoid future deals needing to be brokered at the level of the white house (Oliver Hart and Bengt Homstrom 2016 Nobel). (Though for truly high stakes negotiations this may be unavoidable- i.e. the recent deal for Merck to produce the Johnson & Johnson COVID-19 vaccine (citation).)

Are consortia harmful? Have they become monopolies?

Adam Smith, commonly known as “the father of economics,” once famously said: “People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices” (Wealth of Nations, 1776). Applied to consortia, in this section, we ask if research consortia can potentially be harmful to the broader research community and the public. In addition to the previous examples of Matthew’s effect in consortia, here we highlight its application to research funding: funding can get tied up in large consortia. Once a consortia starts, it’s quite easy to maintain it and justify its continual existence. For example, the National Human Genome Research Institute (NHGRI), a whole branch of the NIH, was created for the HPG and persists today far beyond the completion of the first sequenced human genome. Further, ENCODE which continues today in many iterations and sub-groups, is a direct outgrowth of the HGP (cite). Another example? If not careful, consortia can become analogous to a monopoly.

In the last twenty years, the percentage of zombie companies has increased from nearly non-existent to just under 18.9% of firms in the U.S. (cite). Zombie firms are those that do not make enough profit to pay off their principal debts and only persist because of government bailouts. Zombie firms are not innovating, not profitable, and all together contribute to an unhealthy economy. Their existence stifles innovation of start-ups that both find it difficult to compete against large firms and secure funding in an unhealthy economy (cite to WSJ). Following the money, let us consider bailout money as analogous to grant funding. In Ed Young's piece on the Human Brain Project of the European Commission (citation), he asks aptly: how should funding be allocated, concentrated in large projects or divided among smaller groups? Consortia certainly should not come at the cost of small labs. Generally, innovation and avant-garde ideas can happen more quickly in small groups (Bhattacharya and Packalen, 2020 ?). Famously, a small group doing research on the hot springs of the Yellowstone National Park discovered taq polymerase that could withstand high temperatures and be used in thermocyclers to amplify DNA through PCR (citation). Further, the discovery of CRISPR, a breakthrough Nobel Prize winning tool for precise gene editing across organisms, was also found from a small group working on yogurt research (citation). Research on esoteric ideas is often done in small labs, not consortia. Consortia should not come at the cost of small groups, and funding should be available for both avenues of research. Of course there is a finite amount of research funding and priorities should be balanced between the two. Like zombie companies, biology consortia too are on the rise in the last twenty years. Consortia, unlike zombie firms, are innovating and contributing to an overall healthy research environment, but we must be careful that the funding and proliferation of consortia do not come at the cost of innovation and competition of individual laboratories.

What are the benefits of consortia beyond economic value?

Writing in *Science Magazine* in 1961, before the era of biological consortia, Alvin Weinberg, then director of the Oak Ridge National Laboratory, drew an analogy to the big science projects of the day to historical monuments like the Egyptian pyramids, the cathedrals of the middle ages, and the Palace of Versailles even warning about the link between these projects and the demise of the economies that conceived them. Alvin argued that big science could seriously harm scientific research as a whole, leading to the triple diseases of "journalitis, moneyitis, administratititis" (Weinberg, 1961). In closing he argued that we should focus our efforts on improving human well-being. Today's biological consortia arguably do just that. Half a century later, in 2012, the authors and participants of a review on consortia efforts in immunology, refuted Weinberg, writing: "But, surely, finding a fundamental particle of the Universe or deciphering the human genome has inspirational value at the individual and societal level that transcends any usual science project" (Benoist *et al.*, 2012). In this section, we explore the potential of consortia to, despite the central argument of this piece, transcend economic analyses and provide inspirational value.

As discussed above, consortia science can: democratize tools for science by making them more widely available and affordable, promote open science through collaboration, and lead to the growth of a nation. Even beyond these benefits, as a large-scale effort, consortia have the potential to capture public attention. Some consortia (i.e HGP) have the star power to help shape the public narrative and responsibly foster public trust in science. Trust in science is ultimately a good thing (perhaps we may have better bore the brunt of COVID-19, for example). At times also inherent to their large-scale nature, consortia often represent collaborations across countries and cultures. Not only can this accelerate data access as previously mentioned, these collaborations can bridge across geopolitical boundaries. While not a consortium, during the cold war, for example, the U.S. and the Soviet Union collaborated on bringing a second Polio vaccine to

market. Albert Sabin developed a polio vaccine using attenuated polio virus, but it could not be tested in the U.S. since an earlier vaccine developed by Jonas Salk was in use. In collaboration with Soviet scientists Mikhail Chumakov, Maria Voroshilova, and Anatoli Smorodentsev, Sabin's vaccine was able to be tested in the USSR which had active polio outbreaks, proved efficient, and ultimately used for vaccination of children in both countries (citations). Today, many consortia bridge international boundaries, providing opportunities to scientists from various backgrounds, increasing cultural competency, and ultimately strengthening science itself through the diversity of its participants.

Appendix B

Nature Methods News & Views: Integrative analysis methods to bridge trade-offs in spatial transcriptomics data

This appendix contains the full text of the Nature Methods News & Views piece titled "Integrative analysis methods to bridge trade-offs in spatial transcriptomics data," which was authored jointly by Shaina Lu, Daniel Fürth, and Jesse Gillis. I wrote the original text and contributed to subsequent rounds of substantial editing.

COMPUTATIONAL BIOLOGY

Integrative analysis methods for spatial transcriptomics

Computational methods use different integrative strategies to tackle the challenges of spatially resolved transcriptomics data analysis.

Shaina Lu, Daniel Fürth and Jesse Gillis

Multicellular organisms are defined by the cells that compose them as well as the relationships between those cells, partially captured by cells' spatial organization. Although single-cell transcriptome sequencing (scRNA-seq) has had a transformative impact in characterizing cells as independent elements, many aspects of the cells' relationships are lost with this technique, including spatial distribution. Newly developed tools have focused on assaying the spatial organization of cells in tissues, but there are often trade-offs between spatial resolution and the number of unique RNA transcripts assayed. In this issue of *Nature Methods*, Scalia et al.¹ and Hu et al.² introduce computational tools to integrate spatially resolved transcriptomic data with scRNA-seq and/or histology data to bridge these trade-offs and provide a better understanding of the spatial organization of tissues.

Although focusing on different parts of the analysis process, both SpaGCN² and Tangram¹, the methods of Scalia et al. and Hu et al., respectively, are computational methods for data integration to improve the interpretation of spatial expression (Fig. 1). SpaGCN focuses on incorporating existing histology to identify spatial domains and subsequently identify genes differentially expressed between the spatial clusters. Though Tangram also incorporates aspects of these steps, its principal focus is on providing cross-modality data integration with scRNA-seq data. After this integration, a number of analysis tasks can be accomplished using Tangram, such as imputing additional genes in spatial data that are not transcriptome wide or deconvolving spatial data that are not of cellular resolution into cell-type proportions. The different forms of analysis accomplished by Tangram and SpaGCN are largely complementary.

SpaGCN and Tangram are part of a broader trend toward the development of computational methods for spatial

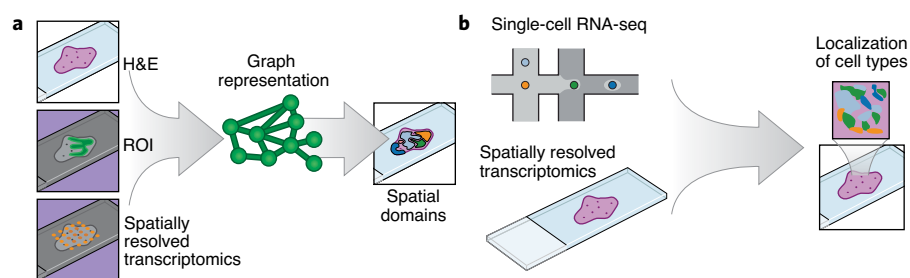


Fig. 1 | Schematic comparison of SpaGCN and Tangram analysis methods for spatially resolved transcriptomics. **a**, SpaGCN integrates histological information, user-defined region of interest (ROI) and spatial transcriptomics into a graph convolutional network (GCN) and performs unsupervised clustering on the graph representation to arrive at a set of spatial domains. H&E, hematoxylin and eosin histochemical staining. **b**, Tangram aligns single-cell data with spatially resolved data to arrive at imputed and deconvolved spatial domains with single-cell-like qualities.

transcriptomics^{3,4}. This development is driven by the increased availability of spatially resolved data and techniques for generating it⁵. SpaGCN is analytically unusual within this cohort for its combined approach to resolving spatial domains and computing differential expression (rather than just one or the other). Like SpaGCN, Tangram uses histology data, but its focus is on aligning any type of single-cell (or single-nucleus) RNA-seq to spatial data packaged with a breadth of methodological tools after integration. Tangram's use as a single-cell and spatial integration tool will be helpful in meeting the popular demand for a straightforward tool to visualize in situ clusters obtained from scRNA-seq^{6,7}. Whereas some earlier tools are specific to one type or class of spatial experiment, both SpaGCN and Tangram can be applied across experimental assays and are meant to be universal tools for the spatial field.

As experimental technologies continue to improve⁵, the gap between high spatial resolution and percentage of the transcriptome assayed continues to shrink⁸. However, until new techniques that promise to cover the whole transcriptome with subcellular resolution are readily

available and accessible, computational data integration is necessary to bridge this gap. Although recent methods are customized for spatial data, the fundamental models are often more general. In essence, information is shared between cells within a dataset in a structured way to minimize noise, and then cells are aligned across datasets. If spatial metadata are available for one of those sets of cells, or the way information is shared between cells is defined by known location, then these data integration methods become spatial data integration methods.

A prominent discussion point in Scalia et al.¹ is the promise of data integration approaches for bringing us closer to a truly multimodality understanding of biology through the creation of large, integrated datasets such as the Human Cell Atlas⁹. Because cellular location is among the most fundamental types of metadata, integration of spatial data is important for large-scale data integration into a common framework. This will allow evaluation of the underlying data and methods, currently a major challenge within the field. As methods improve and reference data emerge, uncovering novel drivers of variability that contribute to disease or other phenotypic

differences should also become possible. Although some phenotypic differences reflect cell-autonomous variability, a substantial fraction is likely emergent from the relationships between cells. Uncovering the logic of how these cell-cell relationships contribute to tissue function is an important avenue opened up by these integrative methods and the data underlying them.

An important area for technical improvement in analysis methods rests on the fact that current assessments are quite qualitative in nature. Although this does not place a direct limit on the efficacy of methods, it does place a limit on our understanding of how best to apply them or improve upon them. Spatial clustering methods or identification of spatial distributions of cell types, for example, are often visualized with microscopy images and are said to be good representations when these computationally defined features match the cytoarchitecture and morphology of the tissue. There are some popular statistical measures, such as those for determining spatial autocorrelation, but these do not capture the performance of all classes of spatial analysis tasks. In addition to the advances in spatial analysis represented by Tangram and SpaGCN, other spatial tools, not detailed here, are also useful. As with any new field, to better understand the pros and cons of the many spatial analysis tools, an independent, rigorous and quantitative benchmarking across spatially resolved transcriptomics analysis tools is needed.

Moving forward, tools such as SpaGCN² and Tangram¹ will be invaluable in establishing spatial regions directly derived from gene expression data, rather than

defined from traditionally agreed anatomical boundaries. Although gene expression need not be the be-all and end-all, it provides a unified and quantitative framework to link activity at the cellular and tissue levels. Boundaries defined from spatial expression will link processes such as cell-cell communication, cell migration and morphogenesis in organ formation. Analysis tools for spatially resolved transcriptomics usually take a data-first approach to understanding biology, sometimes described as ‘unbiased’, but integration with existing biological knowledge to understand causal mechanisms will ultimately require testable hypotheses in combination with high-quality data.

Particularly important for future study are questions relating to evolution and development, as well as their interplay, as modular expansion of spatial domains to create new functions is a repeated theme of both. Evolution and development offer a vast space from which to collect data, with a new class of integration to consider, for which systematic tools such as SpaGCN and Tangram will be essential. Although these tools can capture biological phenomena such as morphological patterns in the brain, clusterings have difficulty in distinguishing between byproducts of evolution and phenotypic traits that are the direct products of selection. Spatial expression across development should provide valuable insight into molecular mechanisms, whereas spatial expression across species helps to capture selection and conservation.

The rapid parallel development of molecular tools available both in spatial genomics⁵ and in lineage tracing and

clonal identification¹⁰ will, together with computational methods like SpaGCN and Tangram, enable a new era of experimental design and discovery. Spatially resolved transcriptomics has the potential to be the revolution of this decade, much as single-cell techniques were for the previous one; these analysis tools will help to realize that potential. □

Shaina Lu , Daniel Fürth  and Jesse Gillis 
Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA.
✉e-mail: JGillis@cshl.edu

Published online: 28 October 2021
<https://doi.org/10.1038/s41592-021-01272-7>

References

- Scalia, G. et al. *Nat. Methods* <https://doi.org/10.1038/s41592-021-01264-7> (2021).
- Hu, J. et al. *Nat. Methods* <https://doi.org/10.1038/s41592-021-01255-8> (2021).
- Svensson, V., Teichmann, S. A. & Stegle, O. *Nat. Methods* **15**, 343–346 (2018).
- Edsgård, D., Johnsson, P. & Sandberg, R. *Nat. Methods* **15**, 339–342 (2018).
- Asp, M., Bergensträhle, J. & Lundeberg, J. *Mol. Cell. Dev. Biol.* **42**, e1900221 (2020).
- La Manno, G. et al. *Nature* **596**, 92–96 (2021).
- Di Bella, D. J. et al. *Nature* **595**, 554–559 (2021).
- Srivatsan, S. R. et al. *Science* **373**, 111–117 (2021).
- Regev, A. et al. *eLife* **6**, e27041 (2017).
- Wagner, D. E. & Klein, A. M. *Nat. Rev. Genet.* **21**, 410–427 (2020).

Acknowledgements

S.L. is supported by the Edward and Martha Gerry Fellowship funded by The William Stamps Farish Fund and the Gladys and Roland Harriman Foundation. D.F. is supported by a NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation. J. G. is supported by NIH grants R01MH113005 and R01LM012736.

Competing interests

The authors declare no competing interests.



ORGANOIDS

Towards spheroid-omics

The MISpheroid knowledgebase records and organizes experimental parameters from thousands of cancer spheroid experiments, revealing heterogeneity and a lack of transparency in key spheroid research reporting practices.

Timothy L. Downing

For more than 40 years, researchers have explored the development of cell culture models that recapitulate biological processes as they occur within three-dimensional (3D) physiological

contexts. However, within the past 10 years, there has been a sharp increase in the rate of spheroid studies published, owing to the valuable insights that these models provide into cancer pathophysiology (including

cell migration and matrix invasion), as well as pharmacological response through drug testing¹. 3D spheroid cultures are established through the aggregation of suspended (non-adherent) cells derived

Appendix C

PLOS Biology: Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain

This appendix contains the full text of the PLOS Biology Methods and Resources Paper titled "Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain," which was authored jointly by Shaina Lu, Cantin Ortiz, Daniel Fürth, Stephan Fischer, Konstantinos Meletis, Anthony Zador, and Jesse Gillis. I co-designed and performed the experiments, wrote the manuscript, made the figures, and co-revised the manuscript.

METHODS AND RESOURCES

Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain

Shaina Lu¹, Cantin Ortiz², Daniel Fürth¹, Stephan Fischer¹, Konstantinos Meletis², Anthony Zador^{1*}, Jesse Gillis^{1*}**1** Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America, **2** Department of Neuroscience, Karolinska Institutet, Solna, Sweden* zador@cshl.edu (AZ); jgillis@cshl.edu (JG)**OPEN ACCESS**

Citation: Lu S, Ortiz C, Fürth D, Fischer S, Meletis K, Zador A, et al. (2021) Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain. *PLoS Biol* 19(7): e3001341. <https://doi.org/10.1371/journal.pbio.3001341>

Academic Editor: Franck Polleux, Columbia University Medical Center, UNITED STATES

Received: January 20, 2021

Accepted: June 29, 2021

Published: July 19, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pbio.3001341>

Copyright: © 2021 Lu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The ABA data is available publicly for download directly from the Allen Brain Atlas website (<http://help.brain-map.org/display/api/Allen%2BBrain%2BAtlas%2BAPI>). ST data (from Ortiz et al., 2020) is available at

Abstract

High-throughput, spatially resolved gene expression techniques are poised to be transformative across biology by overcoming a central limitation in single-cell biology: the lack of information on relationships that organize the cells into the functional groupings characteristic of tissues in complex multicellular organisms. Spatial expression is particularly interesting in the mammalian brain, which has a highly defined structure, strong spatial constraint in its organization, and detailed multimodal phenotypes for cells and ensembles of cells that can be linked to mesoscale properties such as projection patterns, and from there, to circuits generating behavior. However, as with any type of expression data, cross-dataset benchmarking of spatial data is a crucial first step. Here, we assess the replicability, with reference to canonical brain subdivisions, between the Allen Institute's *in situ* hybridization data from the adult mouse brain (Allen Brain Atlas (ABA)) and a similar dataset collected using spatial transcriptomics (ST). With the advent of tractable spatial techniques, for the first time, we are able to benchmark the Allen Institute's whole-brain, whole-transcriptome spatial expression dataset with a second independent dataset that similarly spans the whole brain and transcriptome. We use regularized linear regression (LASSO), linear regression, and correlation-based feature selection in a supervised learning framework to classify expression samples relative to their assayed location. We show that Allen Reference Atlas labels are classifiable using transcription in both data sets, but that performance is higher in the ABA than in ST. Furthermore, models trained in one dataset and tested in the opposite dataset do not reproduce classification performance bidirectionally. While an identifying expression profile can be found for a given brain area, it does not generalize to the opposite dataset. In general, we found that canonical brain area labels are classifiable in gene expression space within dataset and that our observed performance is not merely reflecting physical distance in the brain. However, we also show that cross-platform classification is not robust. Emerging spatial datasets from the mouse brain will allow further characterization of cross-dataset replicability ultimately providing a valuable reference set for understanding the cell biology of the brain.

molecularatlas.org and GEO (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147747>).

Funding: SL is supported by the Edward and Martha Gerry Fellowship funded by The William Stamps Farish Fund and the Gladys and Roland Harriman Foundation. DF is supported by NARSAD Young Investigator Grant from Brain & Behavior Research Foundation. SF is supported by NIH Grant U19MH114821. KM is supported by a grant from the Swedish Research Council (VR project 2018-00608). JG is supported by NIH Grants R01MH113005 and R01LM012736. AZ is supported by NIH Grants 5R01NS073129, 5R01DA036913, RF1MH114132, and U01MH109113, the Brain Research Foundation (BRF-SIA-2014-03), IARPA MICrONS [D16PC0008], Paul Allen Distinguished Investigator Award, Chan Zuckerberg Initiative (2017-0530 ZADOR/ALLEN INST(SVCF) SUB awarded to A.M.Z], and Robert Lourie. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: AZ is a founder and equity owner of Cajal Neuroscience and a member of its scientific advisory board.

Abbreviations: ABA, Allen Brain Atlas; ARA, Allen Reference Atlas; AUROC, area under the receiver operating curve; CFS, correlation-based feature selection; DE, differential expression; ISH, in situ hybridization; k-NN, k-nearest neighbors; LASSO, least absolute shrinkage and selection operator; MWU, Mann-Whitney *U*; PCA, principal component analysis; ST, spatial transcriptomics.

Background

In the last 5 years, there has been an explosion of spatially resolved transcriptomics techniques that have made it possible to easily sequence whole transcriptomes while retaining fine-scale spatial information [1–5]. These new technologies are poised to be transformative across biology [6]. Despite the recent proliferation and improvement of single-cell technologies, these technologies largely depend on tissue dissociation and thus lack information on the spatial origin of sequenced cells. New spatial sequencing tools fill this gap, allowing us to understand the spatial patterning of cell-type specific expression. The stereotyped spatial organization and transcriptional heterogeneity of the brain make it an especially appealing application of these new technologies. Spatial gene expression has the potential to serve as a link between the molecular, mesoscale, and emergent properties of the brain such as gene expression, circuitry, and behavior, respectively [7,8]. This, in turn, could lead to tackling long-standing questions about the brain, such as how gene expression relates to connectivity of neurons or how spatial patterning of expression drives development. Emerging experimental approaches [9–11] and techniques [12–16] have already begun to link multisource information from the mouse brain. However, in order to perform robust multimodality studies, we must first assess replicability within one type of data. Given the potential of spatial transcriptomics (ST) approaches in neuroscience, the early availability of spatial data, and the stereotyped substructure, we use the adult mouse brain as a model system for a cross-platform characterization of spatial data.

Over a decade ago, the first whole-transcriptome, spatially resolved gene expression dataset from the adult mouse brain was collected by the Allen Institute using in situ hybridization (ISH) (Allen Brain Atlas (ABA)) [17,18]. Since its release, this dataset has become a cornerstone for modern neurobiologists who often use it as a first point of reference for gene expression in the mouse brain. The generation of this dataset was a laborious effort requiring many years, the work of many scientists, and many sacrificed mice. The influx of technologies preserving the spatial origin of transcripts presents the opportunity to assess the generalizability of the ABA data for the first time. As the sole reference spatial dataset, benchmarking the ABA data is essential to assess the robustness of the observed gene expression patterns across distinct experiments and technological platforms. In this manuscript, we use “benchmarking” to refer to the assessment of replicability across independent datasets representing different experimental techniques. Obtaining replicable results across gene expression assays is notoriously challenging, so cross-platform, cross-dataset transcriptomics benchmarking has proved crucial since early transcriptome assays in the form of microarrays [19,20].

To address this need for ST and cross-modality robustness in the brain, here we undertook a whole-brain benchmarking of the ABA via linking gene expression and anatomy. We analyzed a spatial gene expression dataset from one adult mouse brain collected using ST [21] (see [Methods](#)) alongside the ABA. ST is a spatially barcoded mRNA capture technique followed by sequencing readout, while the ABA dataset is a collection of single-molecule ISH experiments across the whole transcriptome [1,17]. While benchmarking of the 2 datasets could be done on many scales, we chose to look across brains and across techniques with reference to named brain areas. This approach contains noise associated with the relative biases of each technique (different assays); experimental noise from tissue processing and alignment; biological variability (different brains); and variability from brain area segmentation and naming itself. Despite all these potential sources of noise, our approach combining spatial gene expression with brain area identity allows us to focus on biological conclusions that could be drawn from replicable spatial data. Not readily available with more technical approaches to benchmarking, our approach allowed us to pursue a biological question. We principally ask if canonical, anatomically defined brain areas from the Allen Reference Atlas (ARA) can be assigned using

gene expression alone and, in corollary, how well these assignments replicate across the ABA and ST datasets. We use an interpretable supervised learning framework for classification, where the target values are the ABA brain area labels and the features are the gene expression profiles for samples from across the whole brain (Fig 1A and 1B). We choose to use linear modeling to maintain easily interpretable models that can be related to underlying biology.

Using this approach, we show that ABA labels are classifiable using gene expression, but that performance is higher in the ABA than in ST. We further demonstrate that models trained in one dataset and tested in the opposite dataset do not reproduce classification performance bidirectionally. We then identify potential biological explanations for the difference in cross-dataset performance in classifying brain areas. Finally, we found that although an identifying gene expression profile can always be found for a given brain area, it does not generalize to the opposite dataset. In summary, within each dataset, canonical brain area labels were classifiable and meaningful in gene expression space, but replicability across these 2 very different assays of gene expression was not robust.

Results and discussion

Allen Reference Atlas brain areas are classifiable using gene expression alone

With the advent of new high-throughput capture technologies for ST, we present, as is necessary for all new biological assays, a cross-technology assessment of generalizability in a well-characterized model system: the adult mouse brain. These new technologies allow, for the first time, the cross-platform assessment of canonical, atlas brain area subdivisions relative to gene expression at a whole-brain scale. Traditionally, parcellation of the mouse brain has depended on anatomical landmarks and cytoarchitecture, at times, including interregion connectivity and molecular properties [17,22,23]. By enabling the relatively rapid and high-throughput collection of spatially resolved, whole-transcriptome data in the adult mouse brain, these new spatial assays pave the way for a multimodality assessment of canonical brain area labels. Specifically, in the present work, we ask if brain areas from the ABA [17] are classifiable using 2 spatial gene expression datasets: the Allen Institute's own ISH data [17,18] and a second dataset collected using ST [1,21] (Fig 1A and 1B). After filtering, the ABA consists of 62,527 voxels (rows) with expression from 19,934 unique genes (columns) mapping to 569 nonoverlapping brain area labels, and the ST consists of 30,780 spots (rows) with 16,557 genes (columns) mapping to 461 brain area labels (see Methods for details). The ABA dataset consists of a minimum of roughly 3,260 brains, while the ST dataset is collected from 3 mice (17,21) (see Methods). Comparing accuracy in classification of ABA brain areas across 2 technological platforms and datasets allows us to draw conclusions about spatial expression that are more likely to be biological and generalizable than subject to the technical biases of any one dataset.

To determine if we could more generally determine canonical brain areas from spatial gene expression, we first asked if we could do so within each of the 2 datasets independently. Given the known high correlation structure of gene expression [24], we hypothesized that we could determine the brain area of origin of a gene expression sample using only a subset of the total genes. Fitting these criteria, we chose least absolute shrinkage and selection operator, or LASSO regression [25]. LASSO is a regularized linear regression model that minimizes the L1 norm of the coefficients (i.e., the sum of the absolute values of the coefficients). LASSO typically drives most coefficients toward zero and thus leaves few genes contributing to the final model; LASSO in effect picks “marker genes” of spatial expression in the brain. We use LASSO in a supervised learning framework with a random 50/50 train–test split for two-class classification of all pairwise brain areas successively (Fig 1C) (see Methods). The brain areas included

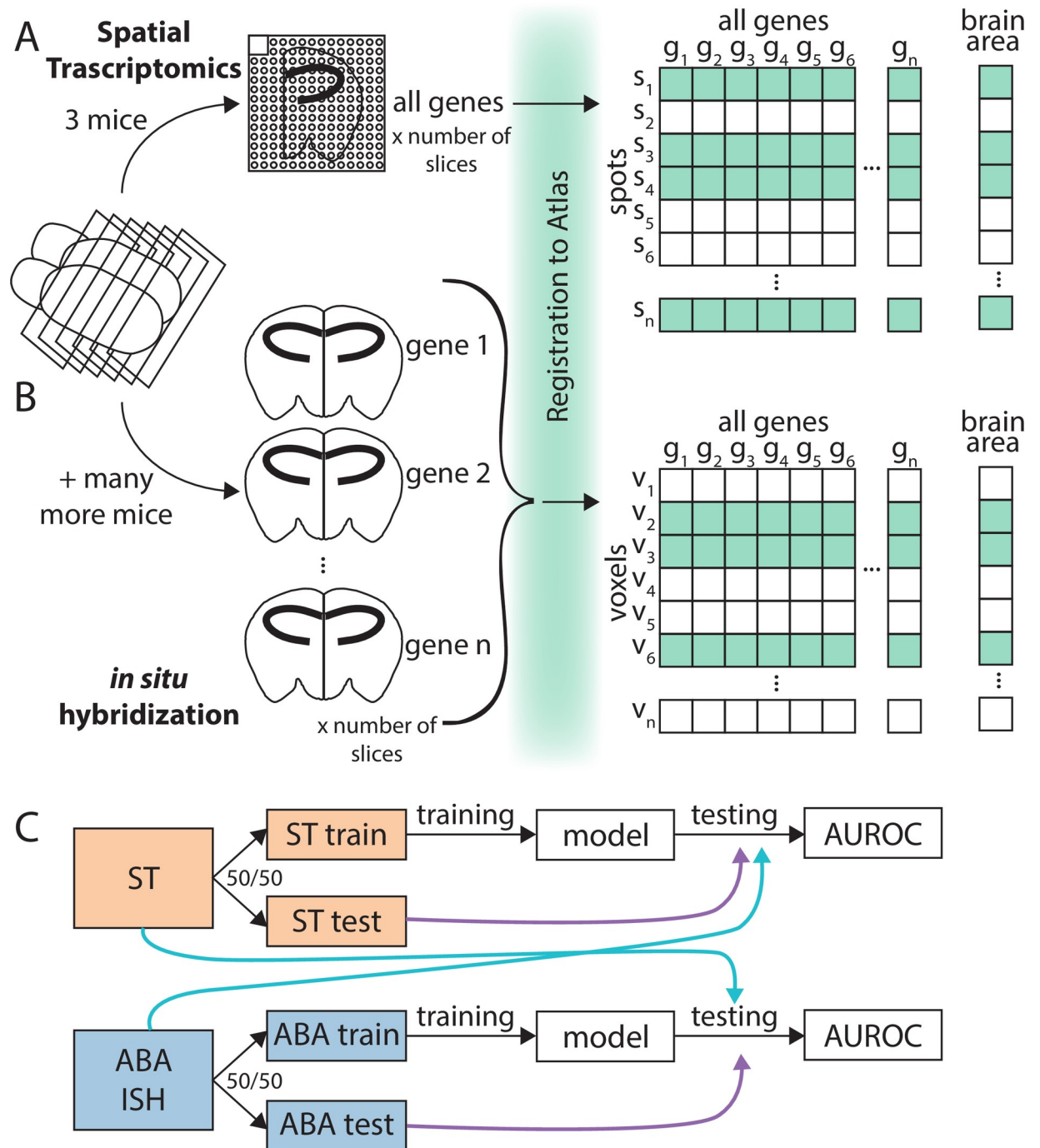


Fig 1. Collection and processing of spatial gene expression datasets. (A) Schematic depicting workflow of collecting whole-brain spatial gene expression using ST. Illustration depicts sectioning of mouse brain, tissue from one hemisphere on one ST slide, registration to ARA, and a layout of the collected data. (B) Schematic depicting workflow of collecting Allen Institute’s whole-brain spatial gene expression using ISH (ABA). Illustration depicts similar workflow to (A), but instead of ST capturing all genes in one (3 for this dataset) brain, there were many more mice used to collect the whole-transcriptome dataset since each brain tissue slice can only be used to probe one gene. (C) Schematic illustrating classification schema. The ST dataset from (A) (orange) and ABA dataset from (B) (blue) were split into 50/50 train/test folds. The training fold was used for model building and the test fold for evaluating the trained model within dataset (purple arrow). Later analysis also applied models trained using the train fold of one dataset to the opposite dataset for testing (light blue arrow). ABA, Allen Brain Atlas; ARA, Allen Reference Atlas; AUROC, area under the receiver operating curve; ISH, in situ hybridization; ST, spatial transcriptomics.

<https://doi.org/10.1371/journal.pbio.3001341.g001>

here are nonoverlapping and are the smallest brain areas present in the ARA naming hierarchy. We subsequently refer to these areas as leaf brain areas since they form the leaves of the tree-based representation of the ARA-named brain areas [17]. The performance of the test set classification is reported using the area under the receiver operating curve (AUROC). The AUROC can be thought of as the probability of correctly predicting a given brain region from its gene expression in a comparison with an outgroup (here, a different brain region) and is calculated by taking the predictions from the trained LASSO model and evaluating their correspondence with the known labels in the test fold (see [Methods](#)). For example, if ranking the samples by the LASSO predictions separates the samples from the 2 classes perfectly without being interspersed, we would get perfect classification with an AUROC of 1, while a score of 0.5 is random. More generally, in this manuscript, we say a brain area pair is classifiable with respect to each other to indicate a high performance in classification with an AUROC greater than 0.5 and generally closer to 1.

After preliminary filtering (see [Methods](#)), we use this approach in both the ST and ABA to classify all the leaf brain areas against each of the others (461 ST areas; 560 ABA areas) ([Fig 1C](#); see [Methods](#)). ARA leaf brain areas are classifiable using LASSO ($\lambda = 0.1$) from all other leaf brain areas using only gene expression data from (1) the ABA (mean AUROC = 0.996) ([Fig 2A](#), [S1A Fig](#)) and from (2) the ST (mean AUROC = 0.883) ([Fig 2B](#), [S1B Fig](#)). These results are consistent across an additional, independent train/test fold split for both datasets (ABA mean AUROC = 0.996, correlation to first split, $\rho = 0.732$; ST mean AUROC = 0.882, correlation to first split, $\rho = 0.860$) ([S1C–S1F Fig](#)). As expected, performance falls to chance when brain area labels are permuted as a control (ABA mean AUROC = 0.510; ST mean AUROC = 0.501) ([S2A–S2D Fig](#)). Together, these results indicate that there is a set of genes whose expression level can be used to identify it and suggests that canonical brain area labels do reflect spatial patterning of gene expression assayed in both the ABA and ST datasets.

Since our task can be conceived as a multiclass classification problem, we asked if brain area classification performance could be improved using a true multiclass classifier. To test this question, we used the *k*-nearest neighbors (*k*-NN) algorithm, which simply assigns the class identity of a test sample based on the majority class label (brain area) of its *k* closest neighbors in feature (here, expression) space. Using *k*-NN ($k = 5$), classification of leaf brain areas fell in ABA (mean AUROC = 0.695; [S2E Fig](#)) and ST (mean AUROC = 0.508; [S2F Fig](#)) (see [Methods](#)). Given the lack of increase in performance and the preferability of our biologically interpretable approach, we choose to continue most analyses using LASSO.

We next asked if single-gene marker selection strategies could outperform LASSO. Highlighting specific brain areas where such markers are known, we looked at classifying the CA2 of the hippocampus and arcuate hypothalamic nucleus with *Amigo2* and *Pomc*, respectively [26–28]. Following long-standing anatomical divisions of the mouse brain, the hippocampal subregions were redefined in the mid-2000s using differences in gene expression [29,30]. Follow-up to the early redefinitions found that while not exclusively expressed in the CA2, *Amigo 2* showed high expression levels in the CA2 [28]. Indeed, in the CA2 of the hippocampus, *Amigo2* performs better than any other single gene in the ABA (*Amigo2* ABA AUROC = 0.920) and ST datasets (*Amigo2* ST AUROC = 0.612) ([S3A Fig](#)). However, classification of the CA2 using *Amigo 2* is still outperformed by the average performance of genes selected by LASSO. One of the major neuronal populations of the arcuate hypothalamic nucleus are the POMC-expressing neurons, shown to have a role in food intake and metabolism [27]. In the arcuate hypothalamic nucleus, *Pomc* performance in the ABA (*Pomc* ABA AUROC = 0.993) and ST (*Pomc* ST AUROC = 0.910) is better than most other single genes and comparable or less than the average LASSO performance for each dataset ([S3B Fig](#)). Given

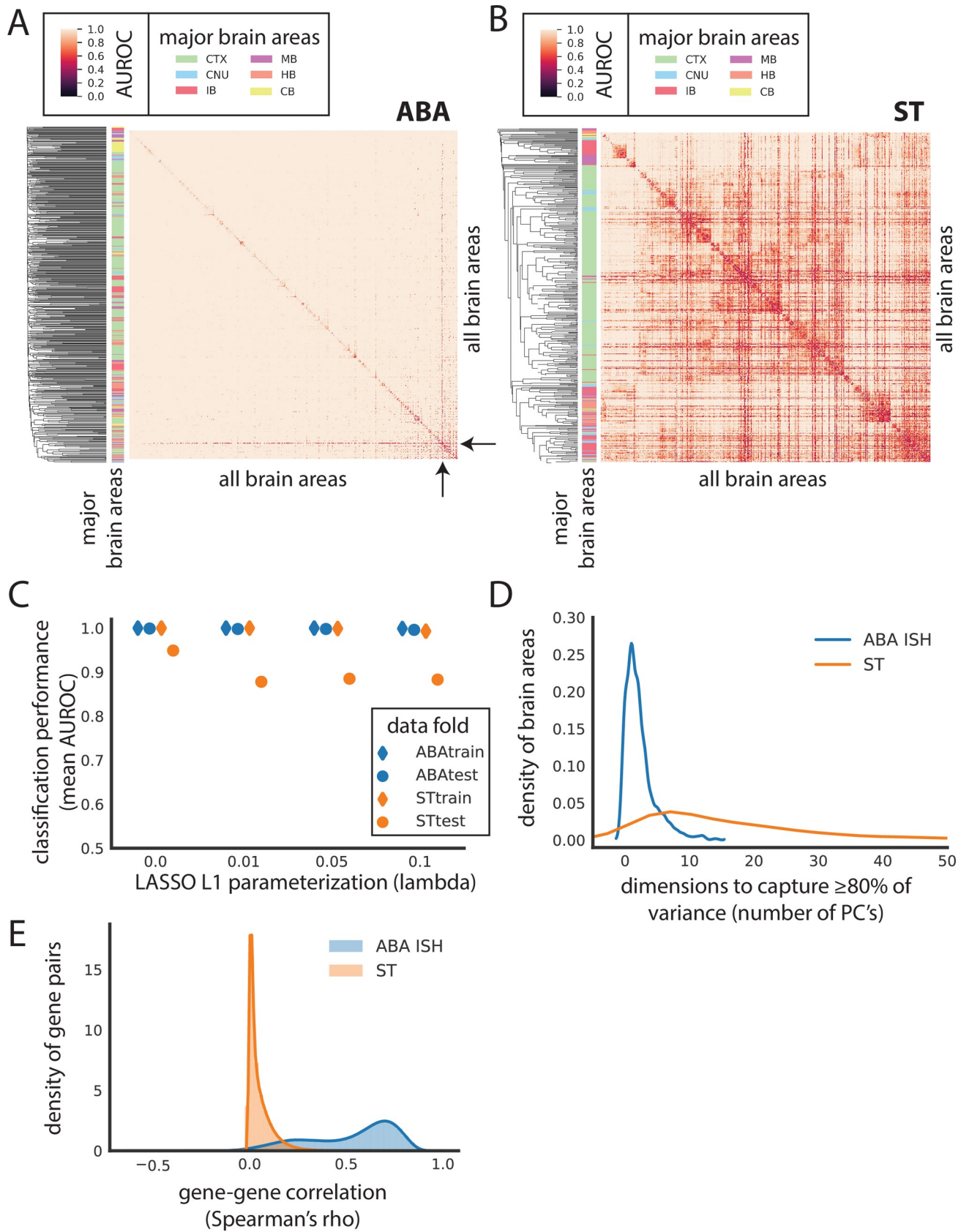


Fig 2. Canonical brain areas are classifiable using gene expression alone in the ABA and ST datasets. Heat map of AUROC for classifying leaf brain areas from all other leaf brain areas in (A) ABA and (B) ST using LASSO ($\lambda = 0.1$). Dendrograms on the far left side represent clustering of leaf brain areas based on the inverse of AUROC; areas with an AUROC near 0.5 get clustered together, while areas with an AUROC near 1 are further apart. Color bar on the left represents the major brain structure that the leaf brain area is grouped under. These areas include CTX, MB, CB, CNU, HB, and IB. (C) Average AUROC (y-axis) of classifying all brain areas from all other brain areas using LASSO across various values of lambda (x-axis): 0, 0.01, 0.05, and 0.1 for ABA train (blue diamond), ABA test (blue dot), ST train (orange diamond), and ST test (orange dot). (D) Number of principal components to capture at least 80% of variance of genes in each of the leaf brain areas after applying PCA to ABA (blue) and ST (orange). ABA brain areas that are larger than ST are randomly down-sampled to have the same number of samples as ST prior to applying PCA. (E) Gene–gene correlations calculated as Spearman’s rho between all pairwise genes across the whole dataset for both the ABA (blue) and ST (orange) independently. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; CB, cerebellum; CNU, striatum and pallidum; CTX, cortex; HB, hindbrain; IB, thalamus and hypothalamus; ISH, in situ hybridization; LASSO, least absolute shrinkage and selection operator; MB, midbrain; PCA, principal component analysis; ST, spatial transcriptomics.

<https://doi.org/10.1371/journal.pbio.3001341.g002>

the comparable performance and, more importantly, since there are not such known markers for most brain areas, we again turned our attention to using LASSO for classifying brain areas.

Notably, performance using LASSO in the ABA is nearly perfect. That the classification in the ABA performs so well is striking, especially considering the potential loss of ISH-level resolution in the voxel representation of the ABA. For the median-performing pair of brain areas in ABA (median AUROC = 1), there is a threshold in classification that can be drawn where all instances of one class can be correctly predicted without any false positives (precision = 1). In contrast, in the ST, no such threshold can be found for the median-performing (median AUROC = 0.959) brain areas (average precision = 0.846) (see [Methods](#)). Further, performance in the ABA is consistently higher than the ST across various parameterizations of LASSO ([Fig 2C](#)) (see [Methods](#)). Despite the comparatively lower performance in the ST, clustering brain areas by AUROC shows brain areas belonging to the same major anatomical region grouping together ([Fig 2B](#)) (see [Methods](#)). For example, most brain areas belonging to the cortex group together in the middle of the heat map (green bar on left) with a few interspersed areas. This grouping suggests that patterns of expression track with broad anatomical labels. Examining the relative expression of genes that are assayed in both datasets, we see that ranked mean expression is comparable across the 2 datasets (Spearman’s $\rho = 0.599$) ([S3C Fig](#)), suggesting that the observed difference in performance is not due to poorly detected genes being well detected in the opposite dataset or vice versa.

Observing the nearly perfect performance in the ABA, we next hypothesized that this dataset may be more low dimensional than suggested by its feature size and may contain many highly correlated features when compared to the ST dataset. We applied principal component analysis (PCA) in each brain area separately by subsetting the data by brain areas, then calculating PCA in each of these subsets independently. Using this approach, we find that on average in individual brain areas, 2 PCs are enough to summarize 80% of the variance per brain area in ABA versus 21 PCs in ST ([Fig 2D](#), [S3D Fig](#)) (see [Methods](#)). In other words, within each brain area in the ABA, many genes are highly coexpressed. Zooming out to the whole brain, using 200 PCs captures nearly 70% of the variance in ABA compared to nearly 20% in ST ([S3E Fig](#)). Further, gene–gene coexpression across the whole dataset is on average higher in the ABA (gene–gene mean Spearman’s $\rho = 0.525$) than in the ST (gene–gene mean Spearman’s $\rho = 0.049$) ([Fig 2E](#)). The perfect performance, low dimensionality on a per brain area basis, and high coexpression all support the idea that although there is meaningful variation in the ABA, it can be captured in few dimensions. In summary, canonical ABA brain areas are classifiable from each other using gene expression alone, but performance is likely inflated in the ABA.

An aside of note is that in the ABA, the one brain area that is consistently lower performing when classified against most other brain areas is the Caudoputamen (mean AUROC = 0.784) ([Fig 2A](#), black arrows). In the ST, the Caudoputamen is not the lowest performing area, but

also has a low mean AUROC (AUROC = 0.619) relative to the other brain areas in ST. In both datasets, the Caudoputamen is the largest leaf brain area composed of the most samples (ABA CP number of voxels = 3,012 versus an average of 85.6 voxels; ST number of spots = 2,051 versus an average of 57 spots). The Caudoputamen is similarly large in other rodent brain atlases, reflecting its lack of cytoarchitectural features [31]. We hypothesized that its relatively larger size could mean that it consists of transcriptomically disparate subsections that are not captured with canonical ARA labeling. Although not an outlier, we do observe that the mean sample correlation for the Caudoputamen in both the ST (mean Pearson's $r = 0.727$) and ABA (mean Pearson's $r = 0.665$) is slightly lower than the mean in either case (ST mean Pearson's $r = 0.783$; ABA mean Pearson's $r = 0.696$) (S4A Fig). More generally, however, we observe that there is no relationship between size and performance across brain regions (S4B and S4C Fig). In addition to being an outlier in terms of size, the Caudoputamen is the dorsal part of the striatum that encompasses many different functional subdivisions evident through the various corticostriatal projections [31]. Together with the low classification performance of the Caudoputamen using gene expression, this reflects the shortcomings of the ARA Caudoputamen label and the likely need to subdivide the Caudoputamen functionally.

Cross-dataset learning of Allen Reference Atlas brain areas

Cross-dataset performance is not bidirectional. Given the low dimensionality and the near-perfect brain area classification performance in the ABA relative to the ST dataset, we hypothesized that the performance of the LASSO models was artificially inflated in the ABA. To explore this hypothesis, we characterized whether LASSO models trained in one dataset would generalize to the opposite dataset (Fig 1C, light blue arrows). For this step, we further filtered for (1) 445 leaf brain areas that were represented with a minimum of 5 samples in each dataset and for (2) 14,299 overlapping genes (see Methods). In this section, we filtered within-dataset analyses to match this set of genes and leaf areas to maintain a parallel evaluation. LASSO-regularized linear models ($\lambda = 0.1$) trained on ST had a similar within-dataset performance (held-out test fold, mean AUROC = 0.884) and cross-dataset performance (ABA, mean AUROC = 0.829) (Fig 3A and 3B), but the reverse is not true. The performance in classifying pairwise leaf brain areas using LASSO models trained in the ABA (held-out test fold, mean AUROC = 0.997) falls when testing in the ST (mean AUROC = 0.725) (Fig 3A and 3C). These results are consistent across an additional random train/test split for both (1) ST (within-dataset ST test mean AUROC = 0.884; correlation to first split, $\rho = 0.735$) to ABA (ST to ABA cross-dataset mean AUROC = 0.831; correlation to first split, $\rho = 0.718$) (S5A and S5B Fig) and (2) for ABA (within-dataset ABA test mean AUROC = 0.997; correlation to first split, $\rho = 0.780$) to ST (ABA to ST cross-dataset mean AUROC = 0.722; correlation to first split, $\rho = 0.816$) (S5A and S5C Fig). These results show that the ST dataset is more generalizable to the opposite dataset than the ABA. Additionally, this discrepancy in cross-dataset performance suggests that the high performance within the ABA is driven by a property of that dataset not present in the ST (see Discussion).

Given this difference in cross-dataset performance, we next explored if correcting for batch effects improves cross-dataset classification performance. We treated each of the 2 datasets as a batch. Batches within each dataset are not clear, particularly in the ABA where batches might arise independently for each gene, which are sampled as an individual experiment by design of single-molecule ISH. After batch correction between the datasets (see Methods), there is virtually no difference in the mean AUROC for either cross-dataset comparison (ABA held-out test fold mean AUROC = 0.997; ABA to ST mean AUROC = 0.725; ST held-out test fold mean AUROC = 0.884; ST to ABA mean AUROC = 0.829). Looking at individual brain area pairs,

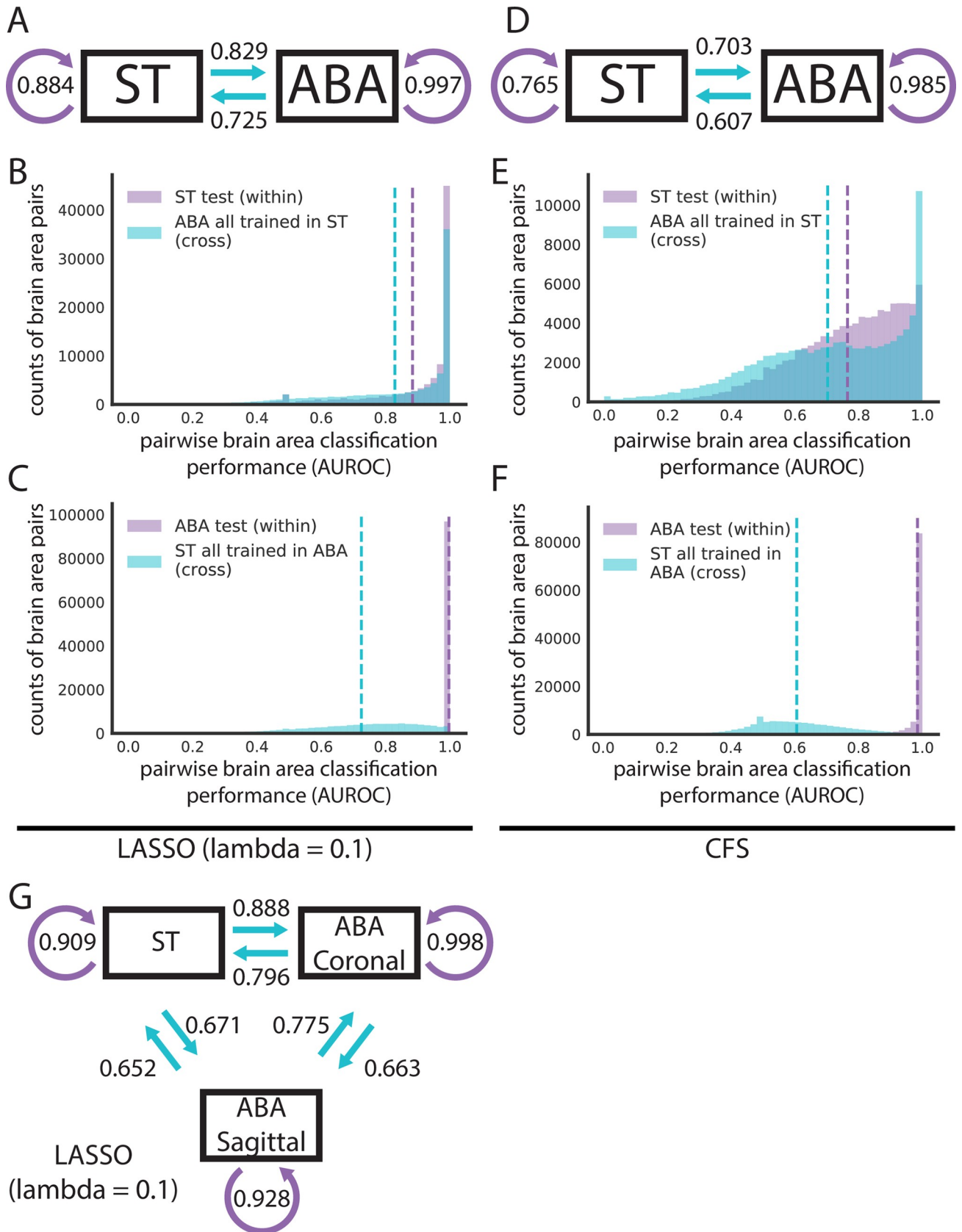


Fig 3. Cross-dataset learning shows that models do not generalize bidirectionally. (A and D) Models trained with overlapping genes and brain areas between ST and ABA datasets are evaluated within dataset on the test fold and across dataset on the entire opposite dataset as illustrated in Fig 1C. Summary diagrams showing mean AUROC for within-dataset test set performance (purple arrow) and cross-dataset performance with models trained in the opposite dataset (light blue arrow) for (A) LASSO ($\lambda = 0.1$) and (D) CFS. Distributions of AUROCs for within- (purple) and cross-dataset (light blue) performance for (B) LASSO ($\lambda = 0.1$) trained in ST, (C) LASSO ($\lambda = 0.1$) trained in ABA, (E) CFS trained in ST, and (F) CFS trained in ABA. In all 4 plots, dashed vertical lines represent the mean of the corresponding colored distribution. (G) Summary diagram showing mean AUROCs using LASSO ($\lambda = 0.1$) for separating out the 2 planes of slicing in the ABA and treating them alongside the ST dataset as 3 different datasets for cross-dataset learning. In all 3 summary diagrams (A, D, G), cross-dataset arrows originate from the dataset that the model is trained in and point to the dataset that those models are tested in. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; CFS, correlation-based feature selection; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.

<https://doi.org/10.1371/journal.pbio.3001341.g003>

there are some minor differences between uncorrected and corrected classification performance with the largest being for the ST within-dataset held out-test fold (mean absolute difference between corrected and uncorrected = 0.001). Hypothesizing that the much-larger ABA dataset could be driving the batch correction and thus showing very little difference between corrected and uncorrected performance, we down-sampled the ABA to have the same sample size as the ST. Filtering, as before, after down-sampling left us with 414 brain areas. Compared to uncorrected performance when filtering for the same brain areas, there are very small differences in the mean AUROCs (see S1 Table). Visualizing the 2 datasets in principal component space suggests that batch correction may not have much effect since there are no obvious global differences relative to one another (S6A–S6F Fig).

We next ask if the high performance seen within ABA that is lost when models built in the ABA are evaluated in the ST is specific to the LASSO method or a more general feature of the data. To assess the data more directly, we used a second simpler method, correlation-based feature selection (CFS). CFS eliminates model building and simply picks features (genes) that are uncorrelated [32] (see Methods). In this way, CFS parallels LASSO, which implicitly picks uncorrelated feature sets when minimizing its L1 regularized cost function that penalizes additional features.

Using CFS, we picked 100 randomly seeded feature sets for pairwise comparisons of leaf brain areas (see Methods; S7A and S7B Fig). We then took the single best-performing feature set from the train set and evaluated its performance on both the held-out test set and cross-dataset. We did this in both directions, training on both ST and ABA as with LASSO above. CFS can accurately classify pairwise leaf brain areas in both the ST (test set mean AUROC = 0.765) and the ABA (test set mean AUROC = 0.985) (Fig 3D–3F). As with LASSO, classification in ABA with CFS is on average better performing than in ST. Again, following a similar trend as LASSO, the difference in mean cross-dataset performance going from the ST test set to the ABA (difference in mean AUROC = 0.052; mean ST to ABA cross-dataset AUROC = 0.703) is smaller than the reverse (difference in mean AUROC = 0.378; mean ABA to ST AUROC = 0.607) (Fig 3D–3F). Altering our analysis approach by averaging the 100 CFS feature sets, we again see a similar pattern in cross-dataset performance (ST to ABA difference in mean AUROC = 0.062; ABA to ST difference in mean AUROC = 0.381) (S7C–S7E Fig). These CFS results indicate that the observed high performance of classification within the ABA and lack of generalization to the ST is not driven by our choice of model. In summary, across both techniques, marker genes can be found to classify pairwise leaf brain areas from each other, but they often do not generalize to the opposite dataset.

The sagittal subset of the ABA is the most distinct. With only 2 datasets, it is impossible to distinguish whether the above lack of bidirectionality in cross-dataset learning is driven by (1) the ST being more generalizable or (2) a lack of information in ST that is critical to the high classification performance within ABA. To begin to address this, we took advantage of the separability of the ABA dataset into 2 distinct datasets: coronal and sagittal. The Allen Institute

collected duplicates of many genes; roughly 4,000 genes were collected across both the coronal and sagittal planes of slicing. With these 2 datasets alongside the ST, we further filtered for 3,737 overlapping genes across the same 445 leaf brain areas (see [Methods](#)) and computed all pairwise combinations of cross-dataset learning. Notably, using LASSO ($\lambda = 0.1$), training on ST outperforms either plane of ABA in cross-dataset predictions: (1) ST to ABA coronal (mean AUROC = 0.888) performs better than ABA sagittal to ABA coronal (mean AUROC = 0.775); and (2) ST to ABA sagittal (mean AUROC = 0.671) performs better than ABA coronal to ABA sagittal (mean AUROC = 0.663) ([Fig 3G](#)). Further, the performance of models trained in ABA coronal to ABA sagittal (mean AUROC = 0.663) and ST to ABA sagittal (mean AUROC = 0.671) is lower than that of ABA coronal and ST to each other (ST to ABA coronal mean AUROC = 0.888; ABA coronal to ST mean AUROC = 0.796) ([Fig 3G](#)). This shows that the ABA coronal and ST are able to generalize to each other better than to the ABA sagittal. Across parametrizations of our model, the sagittal subset of the ABA continues to be the most distinct of the 3 datasets with the least generalizability ([S8A and S8B Fig](#)). To evaluate whether our selection of λ had a significant impact on these findings, we looked at a subset of brain areas with larger sample sizes (minimum of 100) to allow dynamic LASSO hyperparameter fitting and compared it with a fixed hyperparameter ($\lambda = 0.1$) in the same brain areas. This showed that performance was very similar between the two ([S8C and S8D Fig](#)) (see [Methods](#)).

The relative distinctness of the ABA sagittal dataset could be driven by its sparsity—consisting of zeros for more than half of the dataset (53.9%) compared to only 7.5% zeros in the coronal subset. LASSO is able to find a robust set of marker genes within the ABA sagittal that does not reflect the best possible set of genes in the less sparse ABA coronal and ST. While the coronal subset of the ABA was curated for genes showing spatial patterning [[17](#)], the subset of the sagittal genes in this analysis contains only those also present in the coronal set. So, the lack of generalizability of the sagittal subset is particularly suggestive of technical experimental or downstream processing issues rather than the absence of spatial patterning in the genes themselves.

Distance in semantic space, but not physical space, provides a potential explanation for cross-dataset performance

Since the ARA brain areas are organized into a hierarchical tree-like structure based on biology [[17](#)], we hypothesized that the semantic distance of any 2 pairwise brain areas in this tree could provide an explanation for the cross-dataset performance of classifying samples from the same 2 areas. To investigate this, we used the path length of traversing this tree to get from one brain area to the second area as the measure of distance in the tree (see [Methods](#)). For the performance of classifying brain areas in both the ST and ABA when trained in the opposite dataset (LASSO, $\lambda = 0.1$), we see an increase in performance (ST to ABA mean AUROC = 0.690 increases to mean AUROC = 0.912; ABA to ST mean AUROC = 0.655 increases to mean AUROC = 0.756) as the semantic distance increases from the minimum value of 2 to the maximum of 15 ([Fig 4A and 4B](#)). As expected, the corresponding increase in performance and semantic distance holds across parameterizations of our linear model ([S9A–S9D Fig](#)). A high AUROC here indicates that the 2 brain areas are transcriptionally distinct, while an AUROC near 0.5 indicates that they are similar. So, this result implies that distance in semantic space defined by the ARA reflects distance in expression space. This suggests that differences in classification performance are likelier to reflect real differences in gene expression between brain areas and not just large-scale gradients of expression present in the brain [[33](#)].

To further understand the relationship between performance and semantic distance, we next investigated pairs of brain areas with extreme AUROCs at the minimum and maximum

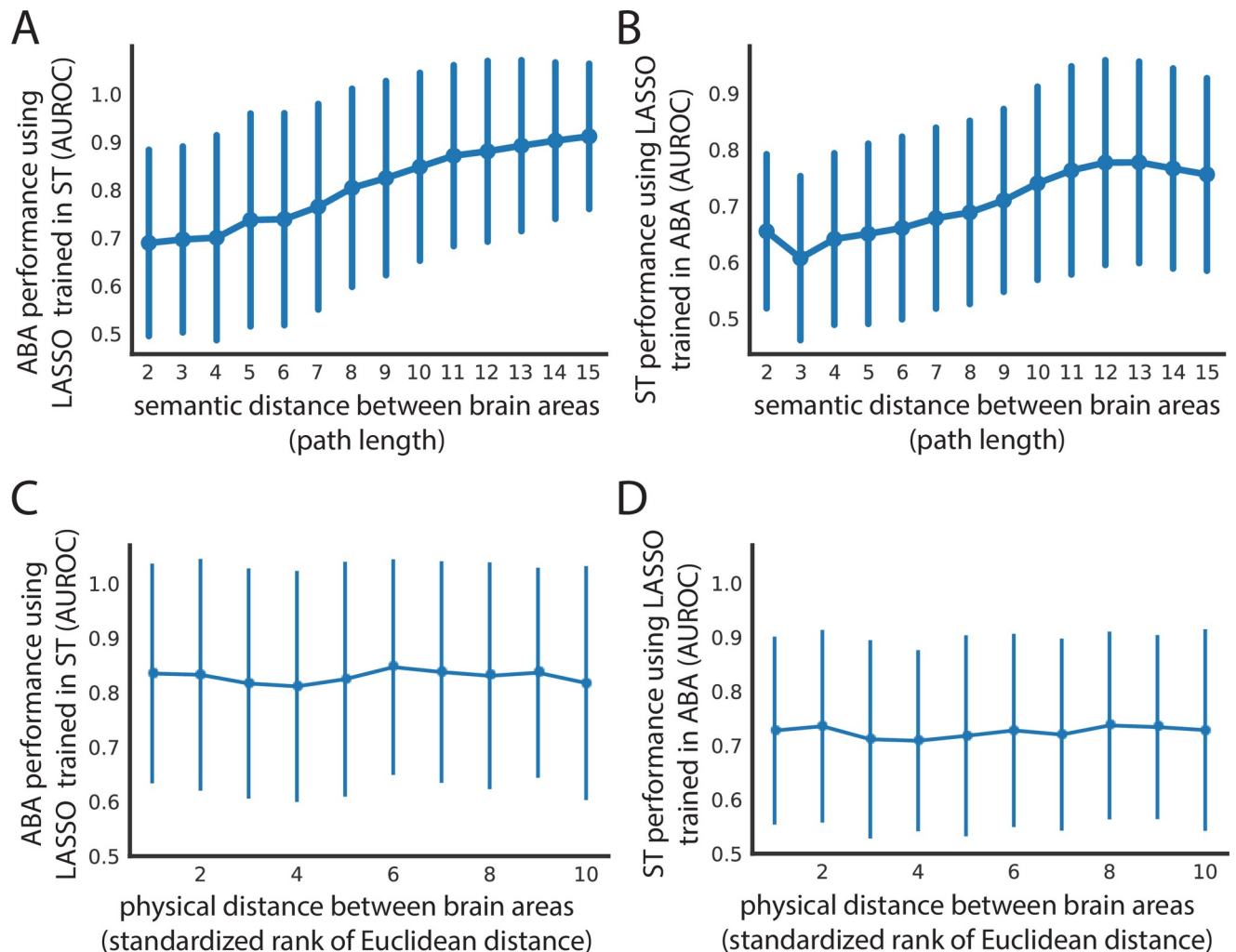


Fig 4. Spatial expression patterns reflect distance in semantic space, but not physical distance in the brain. Cross-dataset AUROCs (x-axis) of classifying all leaf brain areas from all other leaf brain areas for (A) ABA using LASSO ($\lambda = 0.1$) trained in ST and (B) ST using LASSO ($\lambda = 0.1$) trained in ABA as a function of path length (x-axis) in the ARA naming hierarchy between the 2 brain areas being classified. The same AUROCs (y-axis) from (A) and (B) shown in (C) and (D), respectively, as a function of minimum Euclidean distance between the 2 brain areas in the ARA (x-axis). Euclidean distance on the x-axis is binned into deciles for visualization. All 4 plots show mean AUROCs (points) with standard deviation (vertical bars). ABA, Allen Brain Atlas; ARA, Allen Reference Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.

<https://doi.org/10.1371/journal.pbio.3001341.g004>

semantic distances. We were especially interested in this, given the distribution of AUROCs for each distance (Fig 4A and 4B). Similarly, at the smallest semantic distance of 2, in both ABA and ST trained in the opposite dataset, there is a spread in classification performance (S2 and S3 Tables). In both datasets, these brain area pairs involve different cortical layers of the same cortical area. The ARA hierarchy is organized such that within one cortical area, all the layers will have a semantic distance of 2 between each other. So, a pair of brain areas with a high AUROC and semantic distance of 2 often involves 2 nonneighboring layers of a cortical area (i.e., primary auditory cortex layer 6b and layer 4 in ST trained in ABA) (S3 Table). This trend is in line with our expectation as cortical layers are known to have distinct expression profiles driven in part by distinct cell types [34–37]. Alternatively, a pair of brain areas with an AUROC near 0.5 and a semantic distance of 2 can involve 2 neighboring layers of a cortical area (i.e., primary visual area layer 6a and layer 6b in ST trained in ABA) (S3 Table). This, too,

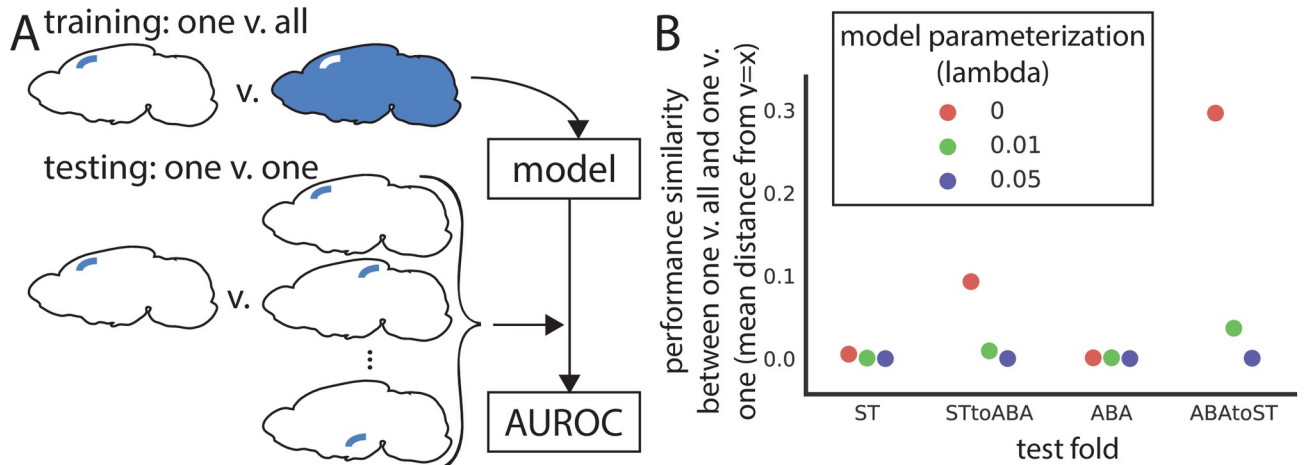
is not surprising because, despite distinctness in cortical layer expression, we expect some overlap between physically neighboring areas in terms of expression profiles due to errors introduced in sampling and in registration to the reference atlas. Together, these examples illustrate one way in which semantic distance is not synonymous to physical distance.

Since semantic distance does not perfectly capture the actual distance between brain areas, we next looked at classification performance as a function of physical distance directly. Specifically, we asked: Is performance in classifying pairwise leaf brain areas cross-dataset being driven by physical proximity/distance alone? Cross-dataset performance was examined with respect to the minimum Euclidean distance between the 2 brain areas in the ARA (see [Methods](#)). There is no trend between physical distance and AUROCs from either cross-dataset assessment using LASSO models trained in the opposite dataset (ABA to ST Pearson's $r = -0.026$; ST to ABA Pearson's $r = 0.056$) with the mean performance remaining similar at the minimum (ABA to ST mean AUROC = 0.651; ST to ABA mean AUROC = 0.702) and maximum distance (ABA to ST mean AUROC = 0.697, change in AUROC = +0.046; ST to ABA mean AUROC = 0.690, change in AUROC = -0.012) ([Fig 4C and 4D](#)) (see [Methods](#)). Across model parameterizations, there is similarly no relationship between distance and performance ([S9E–S9H Fig](#)). This result, alongside the positive relationship seen between performance and semantic distance, shows that spatial patterning of gene expression captures canonical brain area labels and is not merely composed of differences in large-scale gradients.

Finding a uniquely identifying gene expression profile for individual brain areas

Within one dataset, a gene expression profile can uniquely identify one brain area, but it does not generalize to the opposite dataset. Thus far, we have focused on the classification of leaf brain areas from other leaf brain areas. However, this does not determine if we can uniquely identify a given brain area from the whole brain using gene expression. If possible, this could yield a set of marker genes to identify brain areas at their smallest parcellation for future neuroscience experiments. To tackle this, we trained linear models for one leaf brain area against the rest of the brain (one versus all) and tested that same model's performance in classifying the same leaf brain area against all others (one versus one across all leaf brain areas) ([Fig 5A](#)). Unfortunately, for most leaf brain areas, LASSO fails to fit a model with very light regularization ($\lambda = 0.01$) to classify it against the rest of the brain in both the ST (mean train AUROC = 0.554) and the ABA (mean train AUROC = 0.593) ([S10A–S10D Fig](#)). The few leaf brain areas that are able to be classified from the rest of the brain using LASSO have a nearly identical performance in the one versus all case as in testing against all other leaf brain areas ([Fig 5B, S10A and S10B Fig](#)). At a higher regularization weight ($\lambda = 0.05$), most one versus all models fail to be trained (ST mean train AUROC = 0.501; ABA mean train AUROC = 0.502) ([Fig 5B, S10E–S10H Fig](#)). Failing to find potential marker genes using this approach with regularized LASSO, we turned to unregularized linear regression (i.e., $\lambda = 0$), with the hope to minimally find an identifying expression profile. Using linear regression, performance of models fit in the one versus all case correlates nearly perfectly with the average performance of the same model in one versus one. This nearly identical performance is true in both the ST (mean distance from identity line = 0.005) and ABA datasets (mean distance from identity line = 0.001) ([Fig 5B–5D](#)) (see [Methods](#)). This result demonstrates that within a dataset, we can find an identifying gene expression profile of a brain area that uniquely identifies it.

Since we could robustly identify a gene expression profile to identify a brain area within one dataset, we next asked if these profiles can generalize to the opposite dataset. Using the



linear regression trained in ABA

linear regression trained in ST

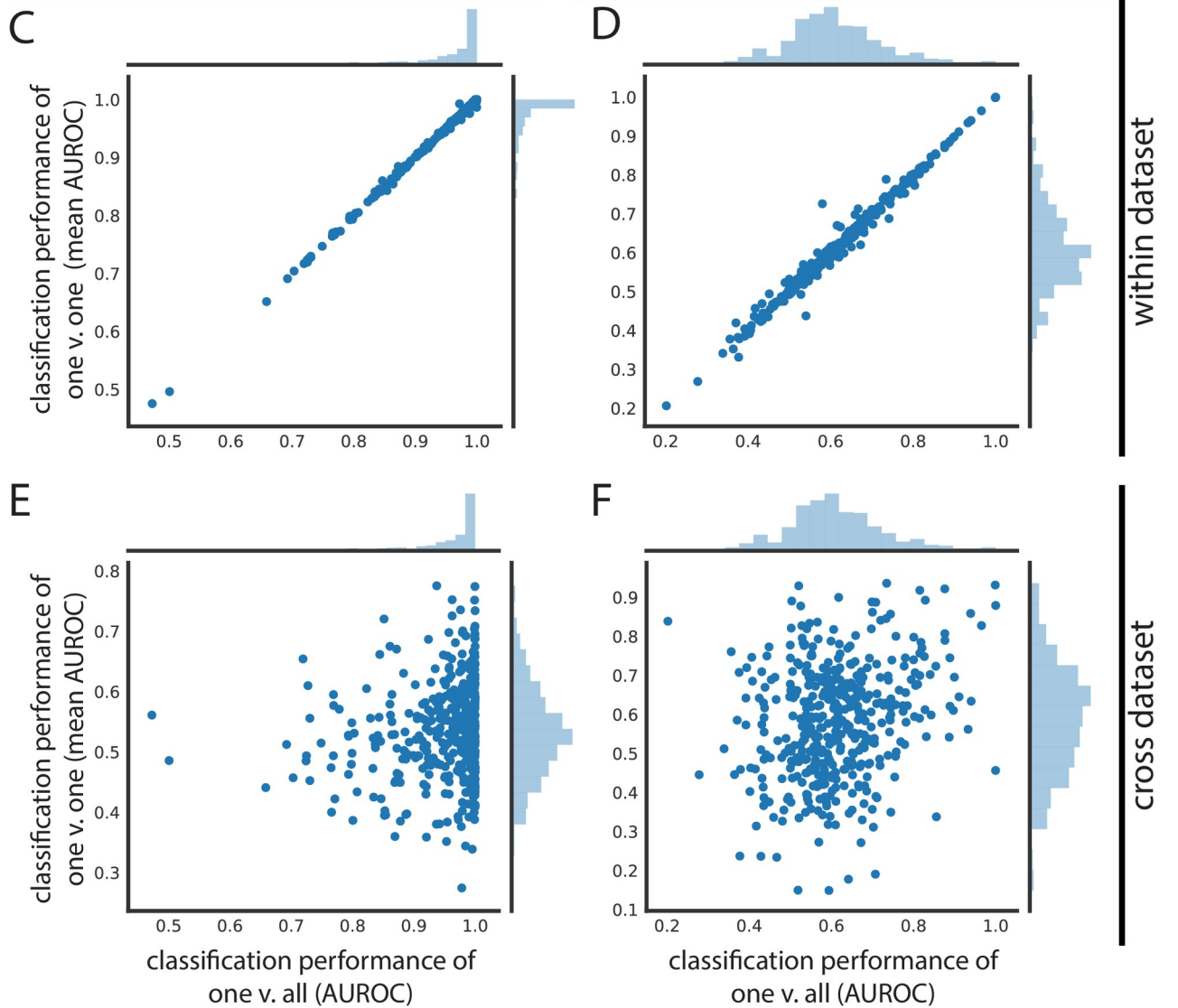


Fig 5. Leaf brain area expression profiles are identifiable within dataset but do not generalize cross-dataset. (A) Schematic depicting training and testing schema for panels in this figure. Models are trained to classify one leaf brain area against the rest of the brain (one vs. all) and then used to test classification of that brain area against all other leaf brain areas (one vs. one) within and across dataset. (B) Performance similarity between one vs. all and one vs. one reported as the mean absolute value of distance from identity line for scatter plots of testing in one vs. all against one vs. one. Performance similarity shown for within ST, ST to ABA, within ABA, and ABA to ST across linear regression (red), LASSO ($\lambda = 0.01$) (green), and LASSO ($\lambda = 0.05$) (violet). Linear regression one vs. all test set performance (x-axis) vs. average one vs. one performance of the same model (y-axis) in (C) ABA and (D) ST. (E) Assessment of the same ABA one vs. all linear regression model (x-axis) in one vs. one classification in the ST dataset (y-axis). (F) Same as (E), but one vs. all linear regression trained in ST (x-axis) and one vs. one classification of these models in ABA (y-axis). ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.

<https://doi.org/10.1371/journal.pbio.3001341.g005>

same models trained in one versus all in either the ST or ABA, we classified the same brain area against all other brain areas (one versus one) in the second dataset. The one versus all trained linear models ($\lambda = 0$) do not generalize cross-dataset for either ABA to ST (mean distance from identity line = 0.296) or the reverse (ST to ABA mean distance from identity line = 0.093) (Fig 5B, 5E, and 5F). This lack of cross-dataset performance similarly holds for other parameterizations (Fig 5B, S10C, S10D, S10G, and S10H Fig). The identifying gene expression profile of a leaf brain area is not generalizable to a new dataset that is not used in defining that profile. So, while we can uniquely identify a brain area using gene expression within one dataset, that identification profile does not extend to the second dataset.

Conclusions

Across disciplines, benchmarking studies have helped to advance their respective fields and set standards for future research [20,38–40]. Neuroscience is no exception. Given the complexity of the brain, however, the addition of multimodal information is especially desirable. Studying the brain from a variety of perspectives, across data modalities, technology platforms, and experiments, can give a complete, composite understanding of its biology [7]. Coupled with the stereotyped substructure and transcriptional heterogeneity, the adult mammalian brain is the ideal model system to assess new spatial gene expression technologies. With this in mind, we linked 2 modalities to ask: Can we capture canonical, anatomically defined brain areas from the ARA using spatial gene expression alone? And, how well does this replicate across 2 transcriptomic datasets collected using different platforms?

Principally, we showed that ARA brain labels are classifiable using only gene expression but highlighted a lack of generalizability across spatial transcriptomic datasets. Within datasets, we are able to distinguish brain areas from each other with high performance. We were further able to uniquely identify a brain area within dataset; training on one brain area against the entire rest of the brain generalizes to testing on that same brain area against all other leaf brain areas. Notably, within-dataset performance was on average higher in the ABA than in the ST, which led to a lack of cross-dataset generalizability when training in the ABA and testing in ST; this phenomenon was not present in reverse. However, in both cases, there is an observed trend linking an increase in mean cross-dataset performance with increased semantic distance in the ARA brain area label organization. There was no link in performance when compared to physical distance in the brain, suggesting that ARA labels are meaningful in expression space and we are not simply detecting spatial differences in gene expression.

It is important to point out that our benchmarking study at its core only involves 2 independent datasets, although both are extraordinary in scope. Limited to 2 datasets, it is impossible to tell whether one dataset or the other is closer to representing the ground truth. Further, the ABA and ST datasets use 2 fundamentally different techniques: The ABA data report average pixel intensity from ISH, and the ST approach is an RNA capture technique followed by sequencing that reports read counts. In addition, the spatial resolution varies between the 2 datasets. Each sample in the ST is further apart within a plane due to a 200- μm center-to-

center distance for probe spots [1] in comparison to samples in the ABA ISH with $200\ \mu\text{m}^3$ voxels that tile adjacently [17]. Conversely, the ABA ISH has a lower Z resolution or larger gap between slices ($200\ \mu\text{m}$) when compared to the ST (median slicing period of $100\ \mu\text{m}$) [21]. Further, after the collection of the raw expression data, each of the 2 datasets also undergoes a unique registration step to the ARA. The ABA uses an iterative approach that involves registration of the 3D brain volume with interspersed smoothing steps [18], while the ST dataset is registered on a slice-by-slice basis to the nearest representative 2D ARA slice using anatomical landmarks [41]. Beyond registration, there are additional concerns about the stability of the ARA brain area labels since there are inconsistencies with other brain atlases and even across versions of the ARA [42–44]. Brain atlases are an imperfect formalism of brain substructure but are the best systematic representation to test spatial gene expression by biological areas.

Past these technical differences, it is also possible that the lack of strong cross-dataset generalization could represent true biological brain-to-brain variability of individual mice. Slight changes in cellular composition between individuals near borders of brain areas could be responsible for the differences between these areas. Zooming out, the results of this manuscript have many potential implications for neuroscience. First, we observed that brain regions are comprehensively better defined by a combination of many genes as opposed to individual markers. Traditionally, however, the description and/or subsequent experimental identification of brain areas by gene expression, often via specific populations of neurons, usually depended on 1 or 2 marker genes [45,46]. The choice to use single marker genes is usually one of practicality, but as spatial experimental and computational analysis techniques continue to improve, the possibility of using better resolved brain areas defined by a multigene expression profile is within reach [21]. This in turn could inform downstream experimental identification of brain areas. Secondly, our results show that quantitative exploration of the existence of brain regions is timely, just as single-cell data have made quantitative exploration of cell type definitions timely. The definition of brain regions is necessary to study the brain, but existing parcellations are almost certainly not sufficient; it is easy to overfit to rigidly defined areas, glossing over individual differences. This highlights the need for continued development of approaches for validating and integrating spatial data from multiple sources. The ST data are a technology that is timely to integrate with single-cell data and valuable to validate, refine, and discover an iterative neuroanatomy that grows with new data types and sources [47]. Finally, an iterative, multimodal definition of brain areas could aid in the research of cross-species comparisons and disease phenotypes, where the mapping of neuroanatomical landmarks is potentially complex.

As the types and prevalence of spatial gene expression approaches continue to increase [5], whole-brain spatial gene expression datasets will surely follow. By continuing to integrate these emerging datasets, we will be able to perform more robust meta-analyses, giving us a deeper understanding of both spatial gene expression with respect to ARA labels and the replicability of spatial technologies in general. An added benefit of the continued incorporation of additional datasets is that, at some point, differences in experimental platforms and registration approaches will only contribute to the robustness of any biological claims. We believe that continued meta-analysis of spatial gene expression in the adult mouse brain and other biological systems is an important route toward integration of distinct data types—location and expression—to form the beginnings of a robust, multimodal understanding of the mammalian brain and other systems.

Methods

Spatial transcriptomics (ST) data

ST is an array-based approach where a tissue section is placed on a chip containing poly-T RNA probes that the mRNA transcripts present in the tissue can hybridize to [1]. These probes

tile the chip in 100- μm diameter spots and contain barcodes specific to that spot so that RNA sequencing reads can be mapped back to their original grid location. Note that the probe spots are not perfectly adjacent to each other but have a center-to-center distance of 200 μm [1].

Here, we used a previously published spatial gene expression dataset containing 75 coronal slices from one hemisphere of the adult mouse brain across 3 animals [21]. The coronal slices were mapped to the Allen Mouse Brain Reference Atlas using a nonrigid transformation approach [41]. In total, this dataset contains 34,103 ST spots across 23,371 genes [21].

Allen Brain Atlas (ABA) in situ hybridization data

The ABA adult mouse ISH dataset consists of a transcriptome-wide assay of expression in inbred WT mice using single-molecule ISH [17]. To assay the whole transcriptome, many WT mouse brains were sliced into 25- μm thick slices containing 8 interlayered sets for subsequent single-molecule hybridization or for staining to create the reference atlas. This results in a z resolution of 200 μm for each gene. These independent image series are subsequently reconstructed to 3 dimensions and registered to the reference brain atlas in interlayered steps [18]. There are 26,078 series, or experiments, across both coronal and sagittal planes with 19,942 unique genes represented. This suggests that a minimum of roughly 3,260 mice brains were used in this dataset, which does not include series that were unused or used for reference staining. These 3D registered reconstructions are then segmented to 200 μm^3 voxels with an associated brain area label. There are 159,326 voxels, with 62,529 mapping to the brain. Gene expression for each of the assayed genes was quantified in these voxels from the imaged data as energy values, which is defined as the sum of expression pixel intensity divided by the sum of all pixels.

The quantified ISH energy values dataset was downloaded from the ABA website (<http://help.brain-map.org/display/mousebrain/API>) through their API on March 12, 2019.

Allen Institute reference brain ontology, leaf brain areas, and path length

The Allen Institute reference brain atlas has organized brain areas into a hierarchy described by a tree data structure. Leaf brain areas are defined here as brain areas that constitute leaves on the ontology tree, i.e., they have no children. Leaf brain areas represent the most fine-scale parcellation of the brain. Using leaf brain areas circumvents the fact that the depth of the tree representing the hierarchical naming structure of brain areas in the ABA is not uniform. Path length refers to the number of steps required to go from one brain area to another in this tree.

Data filtering and train/test split

The ST data were preprocessed to remove ST spots mapping to ambiguous regions, fiber tracts, or ventricular systems and to remove genes that were expressed in less than 0.1% of samples. This left 30,780 ST spots, or samples, with 16,557 genes. For within-ST analyses, this dataset was further filtered to 461 leaf brain areas that each had a minimum of 5 spots. In all analyses, these spots are subsequently randomly split into train and test sets with a 50/50 split. The train/test split is random but stratified for brain areas so that each fold has roughly 50% of the samples belonging to each brain area. N-fold (here, 2) cross validation was used, and results are reported as a mean across folds.

Similarly, the ABA data were filtered for only voxels mapping to the reference brain and genes with expression in at least 0.1% of samples. This gives 26,008 series across 62,527 voxels, also split as described for ST into 50/50 train and test folds. There are 4,972 genes that are assayed more than once across independent experimental series. Except for the analyses

separating out the 2 planes of the ABA data (detailed below), genes duplicated across series were averaged for each voxel for a total of 19,934 unique genes. For within-ABA analyses, this dataset was further filtered to 560 leaf brain areas that each had a minimum of 5 voxels prior to the train/test split. As with ST, within-ABA training, and testing, n-fold (here, 2) cross validation was used, and results are reported as a mean across folds.

For cross-dataset learning, both datasets were further filtered for 445 leaf brain areas that were represented with a minimum of 5 samples in each dataset. Genes were also filtered for those present in both datasets resulting in 14,299 overlapping genes between the two. This filtered subset was used for cross-dataset test set classification and matched within-dataset test set comparisons. For analyses separating out the 2 ABA planes, a similar mapping process was used to determine overlaps between each of the planes and the ST data. This resulted in 3,737 overlapping genes across the same 445 leaf brain areas. Genes that were duplicated in the ABA dataset with independent imaging series within a plane were averaged.

Area under the receiver operating characteristic (AUROC), clustering using AUROC, and precision

The AUROC is typically thought of as calculating the area under the curve of true positive rate as a function of false positive rate. Here, the area under the empirical ROC curve is calculated analytically since it is both computationally tractable and accurate for a given sample. It is given by

$$AUROC = \sum_i^N \frac{Ranks_i}{N_{Pos} * N_{Neg}} - \frac{N_{Pos} + 1}{2 * N_{Neg}}$$

where ranks are the ranks of each positive label sorted by feature, and N_{Pos} and N_{Neg} are the number of positive and negative labels, respectively. This formula is based on the relationship between the Mann–Whitney U (MWU) statistic and AUROC [48–50]. An AUROC of 0.5 indicates that the task being evaluated is performing at chance, while an AUROC of 1 indicates perfect performance. For within-dataset analysis (Fig 2), any AUROCs of 0 were removed from downstream reporting of distributions and mean AUROCs. Note, this filtering does not alter the reported means to the third decimal place. For within-dataset analyses, AUROCs are reported as the mean across 2-fold cross validation.

Clustering by AUROC is done by converting AUROC to a similarity metric by subtracting 0.5 to center the AUROC values at 0.5 and taking the absolute value. The rationale is that if a classification task performs with an AUROC of 0.5, the 2 classes are so similar that they are not distinguishable so they should be grouped closely.

Here, we calculate precision for the median performing brain area pair given by AUROC for within-dataset analysis. We use a threshold that includes all instances of one class, here, all instances of one brain area. Precision is calculated as:

$$precision = \frac{true\ positives}{(true\ positives + false\ positives)}$$

Note that the AUROC of median performing brain area pairs are calculated from the averaged AUROCs across 2 folds, while the reported precision is the average precision of all median brain area pairs from each fold independently because the reported median AUROC (from the fold averaged AUROCs) does not match to actual brain area pairs in either fold.

LASSO and penalty hyperparameter selection

Least absolute shrinkage and selection operator, or LASSO regression, uses an L1 penalty for fitting the linear regression model [25]. The cost function to minimize is given by:

$$\text{cost function} = \min_{\omega} \frac{1}{2n_{\text{samples}}} \|X\omega - y\|_2^2 + \lambda \|\omega\|_1$$

where X represents the matrix of feature values, y the target values, ω the coefficients, and λ the constant value with which to weight the regularization. The notation $\|\cdot\|_1$ represents the L1 norm. A small α gives little regularization ($\alpha = 0$ is equivalent to regular linear regression). An L1 penalty minimizes the absolute value of coefficients, which has an effect of pushing many coefficients toward zero. This is beneficial for highly correlated data to find an optimal set of features among correlated genes, or features, to use for prediction.

In this manuscript, LASSO models are fit using coordinate descent according to the scikit-learn library [51]. Hyperparameter selection for the penalty weight λ is done through cross validation on a subset of brain areas that have sufficient sample size; we use a cutoff of having greater than 100 samples per brain area, which resulted in 65 areas in ST and 139 areas in ABA. With this subset, we use the StratifiedShuffleSplit function from the scikit-learn library to create 3 folds with a test size of 20% within the 50% train set for each dataset [51]. These folds can overlap with each other but are random and stratified by label. We next use these folds in the GridSearchCV function of scikit-learn to perform hyperparameter selection over λ values of 0.01, 0.05, 0.1, 0.2, 0.5, and 0.9. Note, when returning the best-performing hyperparameter or classification result as an AUROC, GridSearchCV returns the first of ties, which can be misleading with tied performance across many hyperparameters (as is often the case here). In both ST and ABA, most pairwise brain area LASSO models perform best with the smallest given λ of 0.01 (S11A and S11B Fig). Since most brain areas lack the sample size to dynamically fit alpha, we chose a fixed λ value for all brain areas in our brain-wide analyses. Although there is not a clear trend, and keeping the ties or near ties in performance in mind, we use a λ of 0.1 for most of our analyses as larger lambda values tend to only show up for smaller brain areas. The hyperparameter λ used for each analysis is noted throughout the main text. We further perform hyperparameter selection in the cross-dataset case when the planes of ABA are separated out. Using 63 brain areas with greater than 100 samples across all 3 “datasets,” we find that mean AUROCs across pairwise cross-dataset classification was comparable between the dynamically fitted λ and the fixed $\lambda = 0.1$ across brain area pairs (S8C and S8D Fig). For additional details on parameterization, see code scripts (repository availability below).

Linear regression

Linear regression is implemented using scikit-learn with default parameters [51]. Normally, when there are more features than samples, linear regression is underdetermined. In the scikit-learn library, however, instead of returning linear regression as unsolvable, it returns the minimum Euclidean norm. (This is different from Ridge Regression where the L2 norm is incorporated in the cost function.)

K-nearest neighbors (k-NN) algorithm

For applications of k-NN, we used the scikit-learn implementation with default hyperparameters: $k = 5$, weights = “uniform,” algorithm = “auto” [51]. Similar to LASSO, 50% of the data was used as the train set, calculating performance of classification on the other 50% held-out test set. As an output, k-NN gives a 1D vector with the length equal to the number of samples

in the test set. This vector contains a predicted brain area label for each test set sample based on the most highly represented class among each test set sample's k closest neighbors in expression space. To compare this classification result to our other classification approaches, we separated out the 1D vector into a 2D binary matrix with a column for each brain area and rows of the same length as the 1D vector representing samples. Each time a sample is predicted as being a particular brain area, the corresponding row and column are marked with a 1. This matrix is then used to calculate an AUROC for predicting each brain area, or column. The mean AUROC of these brain areas is reported in the manuscript.

Batch correction using pyComBat

For batch correction, we use pyComBat, a recent python-based implementation of ComBat [52–54]. Prior to batch correcting, we normalize each dataset independently using z-scoring. We then run pyComBat on the 2 datasets combined treating each dataset as a batch. We do not include covariates. Corrected data are then parsed into the 2 datasets from the combined matrix for subsequent cross-dataset LASSO analysis.

Assessing dimensionality of data using principal component analysis (PCA)

PCA, as implemented in scikit-learn [51], was used to determine the dimensionality of both datasets. PCA was applied to the genes, or features, for each leaf brain area separately in the ABA and ST datasets. The total number of components to use for dimensionality reduction was set to be equal to the number of samples in each area. ABA areas were down-sampled to have the same number of samples as the corresponding brain area in ST. There are 77 brain areas that exceptionally have fewer samples in ABA than ST, so when down-sampling ABA for these 77 areas, the original sample size was used. Dimensionality of the brain areas was then accessed as the number of PCs needed to explain at least 80% of the variance.

Differential expression and correlation-based feature selection (CFS)

Differential expression (DE) in genes is assayed using MWU. Resulting p -values are not corrected for multiple hypothesis testing since p -values are only used to threshold for very extreme DE genes across brain area comparisons. The uncorrected p -values themselves are not reported as a measure for significant DE.

CFS is a feature selection technique that explicitly picks uncorrelated features [32]. Here, a greedy approach to CFS was implemented. The algorithm first chooses a random seed or gene within the top 500 differentially expressed genes. The next gene is then chosen as the lowest correlated gene to the first one and kept if the set AUROC improves. Subsequent genes are chosen as the least correlated on average to the genes already in the feature set. The algorithm stops once the AUROC is no longer improving. The final set of genes chosen using CFS are then aggregated by equally averaging the values of all chosen genes for each sample. Here, in particular, these feature sets fell in the range of 1 to 29 genes with a median of 2 genes in the ABA and the range of 1 to 47 genes with a median of 4 genes in the ST (S7A and S7B Fig). For more details on exact implementation, see code scripts (repository available below).

For the cross-dataset analysis, when unspecified, 100 feature sets were chosen using this approach, and the single best-performing feature set was then evaluated in both the within-dataset test set and the cross-dataset test set. When indicated accordingly, the 100 CFS feature sets were averaged instead of reporting the performance of the best set alone.

Euclidean distance between 2 brain areas

In addition to brain area labels, the ABA dataset contains x, y, z coordinates for each voxel in the ARA space. So, physical distance between 2 brain areas is calculated as the Euclidean distance between the 2 closest voxels where each voxel belongs to one or the other brain area. Due to the symmetry of brain hemispheres, distance was only calculated in one hemisphere by filtering for voxels with a z-coordinate less than 30. This z-coordinate was visually determined to be the midline of the brain based on 3D visualization of the voxel coordinates. Euclidean distances between brain areas calculated in this manner were used for both the ST and ABA datasets since both are registered to the ARA.

Mean distance from identity line

To assess the replicability of models trained in one brain area versus the rest of the brain (one versus all) in classifying that same brain area against all the others (one versus one), the mean absolute Euclidean distance of a scatter plot of those 2 values from the identity line was calculated. This was done to assess how similar the values in the one versus all case are to the one versus one case for each pair of brain areas. Correlation was found to be lacking because it could yield high correlations when the one versus all and one versus one values were quite different for a given point.

Code

All code used for the analyses described in this manuscript was written in Python 3.7 with supporting packages: jupyterlab 1.0.9, h5py 2.9.0, numpy 1.16.4, scipy 1.3.1, pandas 0.25.0, scikit-learn 0.21.2, matplotlib 3.1.0, and seaborn 0.9.0. All Jupyter notebooks and scripts are available on GitHub at www.github.com/shainalu/spatial_rep.

Supporting information

S1 Fig. Additional visualization and verification of within-dataset LASSO results with a new random train/test split. Histogram of classification performance of LASSO ($\lambda = 0.1$) in (A) ABA test fold and (B) ST. (A) and (B) represent the upper triangular of Fig 2A and Fig 2B, respectively. Black dashed vertical line represents the mean. Heat map of AUROC for classifying leaf brain areas from all other leaf brain areas in (C) ABA and (D) ST using LASSO ($\lambda = 0.1$) using a different random train/test split with a seed = 9 relative to Fig 2A and 2B. Dendrograms on the far left side represent clustering of leaf brain areas based on the inverse of AUROC; areas with an AUROC near 0.5 get clustered together, while areas with an AUROC near 1 are further apart. Color bar on the left represents the major brain structure that the leaf brain area is grouped under. These areas include CTX, MB, CB, CNU, HB, and IB. Histogram of classification performance of LASSO ($\lambda = 0.1$) in (E) ABA test fold and (F) ST. (E) and (F) represent the upper triangular of (C) and (D), respectively. Black dashed vertical line represents the mean. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; CB, cerebellum; CNU, striatum and pallidum; CTX, cortex; HB, hindbrain; IB, thalamus and hypothalamus; LASSO, least absolute shrinkage and selection operator; MB, midbrain; ST, spatial transcriptomics. (TIF)

S2 Fig. LASSO performance with permuted labels falls to chance and k-NN performance. Upper triangular (A) histogram and (C) heat map of AUROC for classifying leaf brain areas from all other leaf brain areas in ABA using LASSO ($\lambda = 0.1$) when brain area labels are randomly permuted. (B) and (D) same as (A) and (C), respectively, but for ST with labels

permuted. For heat maps (C, D), dendrograms on the far left side represent clustering of leaf brain areas based on the inverse of AUROC; areas with an AUROC near 0.5 get clustered together, while areas with an AUROC near 1 are further apart. Color bar on the left represents the major brain structure that the leaf brain area is grouped under. These areas include CTX, MB, CB, CNU, HB, and IB. Distribution of performance (AUROC) of classifying each brain area using k-NN, a multiclass classifier, with default parameters ($k = 5$) for (E) ABA and (F) ST. Mean AUROC is shown in the upper left of each plot. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; CB, cerebellum; CNU, striatum and pallidum; CTX, cortex; HB, hindbrain; IB, thalamus and hypothalamus; k-NN, k-nearest neighbors; LASSO, least absolute shrinkage and selection operator; MB, midbrain; ST, spatial transcriptomics. (TIF)

S3 Fig. Classification using single genes, relative expression across datasets, and PCA. Distribution of classifying (A) CA2 and (B) arcuate hypothalamic nucleus against the rest of the brain using single genes. Distributions of all single genes shown for classification in ABA (blue) and ST (orange). Dashed lines represent the marker gene (A) *Amigo2* or (B) *Pomc* for classification in ABA (blue) or ST (orange). (C) Relative expression between the ST and ABA datasets as a density plot. Expression is plotted as the ranked mean for each gene across all samples. (D) Cumulative explained variance curves for PCA in ST (orange) and ABA (blue). Each curve represents one leaf brain area. The total number of principal components per brain area is equal to the number of samples in that area. ABA areas that had more samples than ST are randomly down-sampled accordingly. For both datasets, the Caudoputamen, the largest region, is removed to allow visualization; full figure shown in inset plot. (E) Cumulative explained variance curves for 200 PCs in the whole ST (orange) and whole ABA (blue) datasets. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; PCA, principal component analysis; ST, spatial transcriptomics. (TIF)

S4 Fig. Sample correlation within brain areas and relationship between size and classification performance. (A) Distribution of sample correlation within each of the leaf brain areas for ABA (blue) and ST (orange). Vertical dashed line represents the mean for the correspondingly colored distribution. Test set AUROC in (B) ST and (C) ABA as a function of the number of samples per brain area. The minimum of the 2 brain areas involved in classification is shown. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; ISH, in situ hybridization; ST, spatial transcriptomics. (TIF)

S5 Fig. Additional verification of cross-dataset LASSO results with a new random train/test split. (A) Models trained with overlapping genes and brain areas between ST and ABA datasets are evaluated within dataset on the test fold and across dataset on the entire opposite dataset as illustrated in Fig 1C. Summary diagrams showing mean AUROC for within-dataset test set performance (purple arrow) and cross-dataset performance with models trained in the opposite dataset (light blue arrow) for (A) LASSO ($\lambda = 0.1$) with a new random train/test split (seed = 9) relative to Fig 3A–3C. Distributions of AUROCs for within- (purple) and cross-dataset (light blue) performance for (B) LASSO ($\lambda = 0.1$) trained in ST with new train/test split and (C) LASSO ($\lambda = 0.1$) trained in ABA with new train/test split. In both plots, dashed vertical lines represent the mean of the corresponding colored distribution. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics. (TIF)

S6 Fig. Visualization of the ABA and ST datasets together in low-dimensional space. Plots showing ABA (blue) and ST (orange) datasets visualized together in low-dimensional space after dimensionality reduction with PCA. (A–C) show ST plotted on top of ABA, while (D–F) show the same PCs with ABA plotted on top of ST. Plots show (A, D) PC1 vs. PC2, (B, E) PC2 vs. PC3, and (C, F) PC1 vs. PC3. ABA, Allen Brain Atlas; PCA, principal component analysis; ST, spatial transcriptomics.

(TIF)

S7 Fig. Feature set sizes for CFS and cross-dataset results for CFS with averaging across feature sets. Distribution of CFS feature set sizes for (A) ABA and (B) ST. Models trained with overlapping genes and brain areas between ST and ABA datasets are evaluated within dataset on the test fold and across dataset on the entire opposite dataset as illustrated in Fig 1C. (C) Summary diagrams showing mean AUROC for within-dataset test set performance (purple arrow) and cross-dataset performance with models trained in the opposite dataset (light blue arrow) using the average of 100 feature sets chosen with CFS. Distributions of AUROCs for within- (purple) and cross-dataset (light blue) performance for 100 averaged CFS picked gene sets trained (D) in ST and (E) in ABA. In both plots, dashed vertical lines represent the mean of the correspondingly colored distribution. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; CFS, correlation-based feature selection; ST, spatial transcriptomics.

(TIF)

S8 Fig. Summary plots for cross-dataset analysis of ST, ABA coronal, and ABA sagittal with various parameterizations. Separating out the 2 planes of slicing in the ABA and treating them alongside the ST dataset as 3 different datasets for cross-dataset learning. Summary diagram showing mean AUROCs using (A) linear regression and (B) LASSO ($\lambda = 0.05$). (C, D) Same cross-dataset analysis as (A, B) but looking only at brain areas with at least 100 samples for (C) fixed $\lambda = 0.1$ and (D) dynamically fitted λ for each brain area pair. In all 4 summary diagrams, cross-dataset arrows originate from the dataset that the model is trained in and point to the dataset that those models are tested in. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.

(TIF)

S9 Fig. Comparison of path length and Euclidean distance to LASSO performance for various parameterizations of LASSO. Cross-dataset AUROCs (x-axis) of classifying all leaf brain areas from all other leaf brain areas for (A) ABA using LASSO ($\lambda = 0.05$) trained in ST, (B) ABA using linear regression trained in ST, (C) ST using LASSO ($\lambda = 0.05$) trained in ABA, and (D) ST using linear regression trained in ABA as a function of path length (x-axis) in the ARA naming hierarchy between the 2 brain areas being classified. The same AUROCs (y-axis) from (A–D) shown in (E–H), respectively, as a function of minimum Euclidean distance between the 2 brain areas in the ARA (x-axis). Euclidean distance on the x-axis is binned into deciles for visualization. All plots show mean AUROCs (points) with standard deviation (vertical bars). ABA, Allen Brain Atlas; ARA, Allen Reference Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.

(TIF)

S10 Fig. One vs. all and one vs. one analysis across various parameterizations. LASSO ($\lambda = 0.01$) one vs. all test set performance (x-axis) vs. average one vs. one performance (y-axis) of the same dataset (A) in ABA and (B) in ST. (C, D) Same as (A and B), but one vs. one performance (y-axis) is accessed in the opposite dataset for (C) train one vs. all in ABA

and test one vs. one in ST and (D) train one vs. all in ST and test one vs. one in ABA. (E–H) Same as (A–D), respectively, but using LASSO ($\lambda = 0.05$). ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.
(TIF)

S11 Fig. Relationship between sample size and LASSO hyperparameter choice. Plots showing the best λ for LASSO when dynamically fit across various possible values as a function of brain area sample size in (A) ABA and (B) ST. ABA, Allen Brain Atlas; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.
(TIF)

S1 Table. Cross-dataset classification performance (mean AUROC) for batch-corrected vs. not-batch corrected ABA and ST datasets. ABA data are randomly down-sampled here to have the same sample size as ST. Note that the not-batch corrected case is not down-sampled, but it is filtered for the brain areas that are included in the batch corrected mean AUROCs after down-sampling. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; ST, spatial transcriptomics.
(PDF)

S2 Table. Examples of brain area pairs from LASSO ($\lambda = 0.1$) trained in ST and tested in ABA with minimum path lengths with low and high AUROCs. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.
(PDF)

S3 Table. Examples of brain area pairs from LASSO ($\lambda = 0.1$) trained in ABA and tested in ST with minimum path lengths with low and high AUROCs. ABA, Allen Brain Atlas; AUROC, area under the receiver operating curve; LASSO, least absolute shrinkage and selection operator; ST, spatial transcriptomics.
(PDF)

Acknowledgments

The authors would like to thank Leon French for providing insight on systematically accessing the Allen Institute data, Manthan Shah for executing this, and Nathan Fox for streamlining datasets used. The authors would also like to thank Jose Fernandez Navarro for mapping the RNA-seq data of the Spatial Transcriptomics dataset. Finally, the authors would like to thank Sara Ballouz, Xiaoyin Chen, Megan Crow, Aki Funamizu, Benjamin Harris, Longwen Huang, Risa Kawaguchi, Elyse Schetty, Colin Stoneking, Jessica Tollkuhn, and Alex Vaughan for useful input and discussion.

Author Contributions

Conceptualization: Anthony Zador, Jesse Gillis.

Data curation: Shaina Lu, Cantin Ortiz, Konstantinos Meletis.

Formal analysis: Shaina Lu, Daniel Fürth, Stephan Fischer, Konstantinos Meletis.

Funding acquisition: Anthony Zador, Jesse Gillis.

Methodology: Shaina Lu, Stephan Fischer, Konstantinos Meletis.

Resources: Anthony Zador, Jesse Gillis.

Supervision: Konstantinos Meletis, Anthony Zador, Jesse Gillis.

Visualization: Shaina Lu.

Writing – original draft: Shaina Lu.

Writing – review & editing: Shaina Lu, Stephan Fischer, Anthony Zador, Jesse Gillis.

References

1. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*. 2016; 353(6294):78–82. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27365449>. <https://doi.org/10.1126/science.aaf2403> PMID: 27365449
2. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*. 2019; 363(6434):1463–7. <https://doi.org/10.1126/science.aaw1219> PMID: 30923225
3. Vickovic S, Eraslan G, Salmén F, Klughammer J, Stenbeck L, Schapiro D, et al. High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods*. 2019; 16(10):987–90. <https://doi.org/10.1038/s41592-019-0548-y> PMID: 31501547
4. Stickels RR, Murray E, Kumar P, Li J, Marshall JL, Di Bella DJ, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nat Biotechnol*. 2020:1–7. <https://doi.org/10.1038/s41587-020-0739-1> PMID: 33288904
5. Asp M, Bergenstråhle J, Lundeberg J. Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration. *BioEssays*. 2020:1900221. <https://doi.org/10.1002/bies.201900221> PMID: 32363691
6. Marx V. Method of the Year: spatially resolved transcriptomics. *Nat Methods*. 2021; 18(1):9–14. <https://doi.org/10.1038/s41592-020-01033-y> PMID: 33408395
7. Lein E, Borm LE, Linnarsson S. The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing. *Science*. 2017; 358(6359):64–9. <https://doi.org/10.1126/science.aan6827> PMID: 28983044
8. Close JL, Long BR, Zeng H. Spatially resolved transcriptomics in neuroscience. *Nat Methods*. 2021; 18(1):23–5. <https://doi.org/10.1038/s41592-020-01040-z> PMID: 33408398
9. Economo MN, Viswanathan S, Tasic B, Bas E, Winnubst J, Menon V, et al. Distinct descending motor cortex pathways and their roles in movement. *Nature*. 2018; 563(7729):79–84. <https://doi.org/10.1038/s41586-018-0642-9> PMID: 30382200
10. Bendesky A, Kwon YM, Lassance JM, Lewarch CL, Yao S, Peterson BK, et al. The genetic basis of parental care evolution in monogamous mice. *Nature*. 2017; 544(7651):434–9. <https://doi.org/10.1038/nature22074> PMID: 28424518
11. Moffitt JR, Bambah-Mukku D, Eichhorn SW, Vaughn E, Shekhar K, Perez JD, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*. 2018; 362(6416):16.
12. Cadwell CR, Palasantza A, Jiang X, Berens P, Deng Q, Yilmaz M, et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat Biotechnol*. 2016; 34(2):199–203. <https://doi.org/10.1038/nbt.3445> PMID: 26689543
13. Hanchate NK, Lee EJ, Ellis A, Kondoh K, Kuang D, Basom R, et al. Connect-seq to superimpose molecular on anatomical neural circuit maps. *Proc Natl Acad Sci U S A*. 2020; 117(8):4375–84. <https://doi.org/10.1073/pnas.1912176117> PMID: 32034095
14. Chen X, Sun YC, Zhan H, Kebschull JM, Fischer S, Matho K, et al. High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing. *Cell*. 2019; 179(3):772–786.e19. <https://doi.org/10.1016/j.cell.2019.09.023> PMID: 31626774
15. Huang L, Kebschull JM, Fürth D, Musall S, Kaufman MT, Churchland AK, et al. BRICseq Bridges Brain-wide Interregional Connectivity to Neural Activity and Gene Expression in Single Animals. *Cell*. 2020; 182(1):177–188.e27. <https://doi.org/10.1016/j.cell.2020.05.029> PMID: 32619423
16. Sun Y-C, Chen X, Fischer S, Lu S, Zhan H, Gillis J, et al. Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections. *Nat Neurosci*. 2021 May; 10:1–13.
17. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature [Internet]*. 2007 Jan 6; 445(7124):168–76. <https://doi.org/10.1038/nature05453> PMID: 17151600

18. Ng L, Pathak SD, Kuan C, Lau C, Dong H, Sodt A, et al. Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. *IEEE/ACM Trans Comput Biol Bioinforma.* 2007; 4(3):382–92. <https://doi.org/10.1109/tcbb.2007.1035> PMID: 17666758
19. Canales RD, Luo Y, Willey JC, Austermilller B, Barbacioru CC, Boysen C, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol.* 2006; 24:1115–22. <https://doi.org/10.1038/nbt1236> PMID: 16964225
20. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol.* 2006; 24(9):1151–61. <https://doi.org/10.1038/nbt1239> PMID: 16964229
21. Ortiz C, Navarro JF, Jurek A, Märtin A, Lundeberg J, Meletis K. Molecular atlas of the adult mouse brain. *Sci Adv.* 2020; 6(26):eabb3446. <https://doi.org/10.1126/sciadv.abb3446> PMID: 32637622
22. Crick F, Jones E. Backwardness of human neuroanatomy. *Nature.* 1993; 361(6408):109–10. <https://doi.org/10.1038/361109a0> PMID: 8421513
23. MacKenzie-Graham A, Lee EF, Dinov ID, Bota M, Shattuck DW, Ruffins S, et al. A multimodal, multidimensional atlas of the C57BL/6J mouse brain. *J Anat.* 2004; 204(2):93–102. <https://doi.org/10.1111/j.1469-7580.2004.00264.x> PMID: 15032916
24. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci.* 1998; 95(25):14863–8. <https://doi.org/10.1073/pnas.95.25.14863> PMID: 9843981
25. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc B.* 1996; 58(1):267–88.
26. Hitti FL, Siegelbaum SA. The hippocampal CA2 region is essential for social memory. *Nature.* 2014; 508(1):88–92. <https://doi.org/10.1038/nature13028> PMID: 24572357
27. Toda C, Santoro A, Kim JD, Diano SPOMC. Neurons: From Birth to Death. *Annu Rev Physiol.* 2017; 79:209–36. <https://doi.org/10.1146/annurev-physiol-022516-034110> PMID: 28192062
28. Laeremans A, Nys J, Luyten W, D'Hooge R, Paulussen M, Arckens L. AMIGO2 mRNA expression in hippocampal CA2 and CA3a. *Brain Struct Funct.* 2013; 218(1):123–30. <https://doi.org/10.1007/s00429-012-0387-4> PMID: 22314660
29. Lein ES, Zhao X, Gage FH. Defining a Molecular Atlas of the Hippocampus Using DNA Microarrays and High-Throughput In Situ Hybridization. *J Neurosci.* 2004; 24(15):3879–89. <https://doi.org/10.1523/JNEUROSCI.4710-03.2004> PMID: 15084669
30. Lein ES, Callaway EM, Albright TD, Gage FH. Redefining the boundaries of the hippocampal CA2 subfield in the mouse using gene expression and 3-dimensional reconstruction. *J Comp Neurol.* 2005; 485(1):1–10. <https://doi.org/10.1002/cne.20426> PMID: 15776443
31. Hintiryan H, Foster NN, Bowman I, Bay M, Song MY, Gou L, et al. The mouse cortico-striatal projectome. *Nat Neurosci.* 2016; 19(8):1100–14. <https://doi.org/10.1038/nn.4332> PMID: 27322419
32. Hall MA. Correlation-based Feature Selection for Machine Learning. The University of Waikato; 1999.
33. Fornito A, Arnatkevi9 A, Fulcher BD. Bridging the Gap between Connectome and Transcriptome. *Trends Cogn Sci.* 2018. <https://doi.org/10.1016/j.tics.2018.10.005> PMID: 30455082
34. Yao Z, Liu H, Xie F, Fischer S, Adkins R, Aldrige A, et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv.* 2020; 5:2020.02.29.970558.
35. Yao Z, Nguyen TN, van Velthoven C, Goldy J, Sedeno-Cortes A, Baftizadeh F, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *bioRxiv.* 2020 31;2020.03.30.015214.
36. Codeluppi S, Borm LE, Zeisel A, La Manno G, van Lunteren JA, Svensson CI, et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods.* 2018 Nov 1; 15(11):932–5. <https://doi.org/10.1038/s41592-018-0175-z> PMID: 30377364
37. Poulin JF, Tasic B, Hjerling-Leffler J, Trimarchi JM, Awatramani R. Disentangling neural cell diversity using single-cell transcriptomics. *Nat Neurosci.* 2016; 19:1131–41. <https://doi.org/10.1038/nn.4366> PMID: 27571192
38. Su Z, Łabaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol.* 2014 Sep 1; 32(9):903–14. <https://doi.org/10.1038/nbt.2957> PMID: 25150838
39. Meta-analysis in basic biology. *Nat Methods.* 2016; 13:959.
40. Robinson MD, Vitek O. Benchmarking comes of age. *Genome Biol.* 2019; 20:205. <https://doi.org/10.1186/s13059-019-1846-5> PMID: 31597556
41. Fürth D, Vaissière T, Tzortzi O, Xuan Y, Märtin A, Lazaridis I, et al. An interactive framework for whole-brain maps at cellular resolution. *Nat Neurosci.* 2018 Jan 4; 21(1):139–49. <https://doi.org/10.1038/s41593-017-0027-7> PMID: 29203898

42. Azimi N, Yadollahikhales G, Argenti JP, Cunningham MG. Discrepancies in stereotaxic coordinate publications and improving precision using an animal-specific atlas. *J Neurosci Methods*. 2017; 284:15–20. <https://doi.org/10.1016/j.jneumeth.2017.03.019> PMID: 28392415
43. Chon U, Vanselow DJ, Cheng KC, Kim Y. Enhanced and unified anatomical labeling for a common mouse brain atlas. *Nat Commun*. 2019 Dec 1; 10(1):1–12. <https://doi.org/10.1038/s41467-018-07882-8> PMID: 30602773
44. Wang Q, Ding SL, Li Y, Royall J, Feng D, Lesnar P, et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell*. 2020; 181(4):936–953.e20. <https://doi.org/10.1016/j.cell.2020.04.007> PMID: 32386544
45. Grange P, Mitra PP. Computational neuroanatomy and gene expression: optimal sets of marker genes for brain regions. 2012.
46. Huang ZJ. Toward a genetic dissection of cortical circuits in the mouse. *Neuron*. 2014; 83:1284–302. <https://doi.org/10.1016/j.neuron.2014.08.041> PMID: 25233312
47. Toga AW, Thompson PM, Mori S, Amunts K, Zilles K. Towards multimodal atlases of the human brain. *Nat Rev Neurosci*. 2006; 7:952–66. <https://doi.org/10.1038/nrn2012> PMID: 17115077
48. Krzanowski WJ, Hand DJ. *ROC. Curves for Continuous Data*. CRC Press Taylor & Francis Group; 2009.
49. Hanley JA, McNeil BJ. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*. 1982:29–36.
50. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves. *Q J R Meteorol Soc*. 2002; 128:2145–66.
51. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011; 12.
52. Behdenna A, Haziza J, Azencott CA, Nordor A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *bioRxiv*; 2020. p. 2020.03.17.995431.
53. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012 Mar; 28(6):882–3. <https://doi.org/10.1093/bioinformatics/bts034> PMID: 22257669
54. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007 Jan 1; 8(1):118–27. <https://doi.org/10.1093/biostatistics/kxj037> PMID: 16632515

Supplementary Table 1

Cross-Dataset Classification Performance (mean AUROC)

	ST to ABA		ABA to ST	
	within test	cross	within test	cross
Batch Corrected	0.897	0.838	0.991	0.715
Not Batch Corrected	0.897	0.839	0.998	0.734

Supplementary Table 2

ST to ABA LASSO, alpha=0.1

auroc file = "STtoABA_ABAall_f1_0p1_051420.csv"

AUROC = 1; path length = 2

Brain Area 1				Brain Area 2			
id	acronym	name	parent	id	acronym	name	parent
9	SSp-tr6a	Primary somatosensory area, trunk, layer 6a	361	1086	SSp-tr4	Primary somatosensory area, trunk, layer 4	361
9	SSp-tr6a	Primary somatosensory area, trunk, layer 6a	361	670	SSp-tr2/3	Primary somatosensory area, trunk, layer 2/3	361
1005	AUDp6b	Primary auditory area, layer 6b	1002	735	AUDp1	Primary auditory area, layer 1	1002
1102	SSp-m6a	Primary somatosensory area, mouth, layer 6a	345	950	SSp-m4	Primary somatosensory area, mouth, layer 4	345
1066	VISam2/3	Anteromedial visual area, layer 2/3	394	1046	VISam6a	Anteromedial visual area, layer 6a	394
308	PTLp6a	Posterior parietal association areas, layer 6a	22	241	PTLp2/3	Posterior parietal association areas, layer 2/3	22
1030	SSp-ll1	Primary somatosensory area, lower limb, layer 1	337	478	SSp-ll6a	Primary somatosensory area, lower limb, layer 6a	337
905	VISal2/3	Anterolateral visual area, layer 2/3	402	601	VISal6a	Anterolateral visual area, layer 6a	402
600	AUDd2/3	Dorsal auditory area, layer 2/3	1011	156	AUDd6a	Dorsal auditory area, layer 6a	1011
251	AUDp2/3	Primary auditory area, layer 2/3	1002	847	AUDp5	Primary auditory area, layer 5	1002
583	CLA	Clastrum	703	780	PA	Posterior amygdalar nucleus	703
1035	SSs4	Supplemental somatosensory area, layer 4	378	862	SSs6a	Supplemental somatosensory area, layer 6a	378
729	TEa6a	Temporal association areas, layer 6a	541	97	TEa1	Temporal association areas, layer 1	541
478	SSp-ll6a	Primary somatosensory area, lower limb, layer 6a	337	113	SSp-ll2/3	Primary somatosensory area, lower limb, layer 2/3	337
816	AUDp4	Primary auditory area, layer 4	1002	954	AUDp6a	Primary auditory area, layer 6a	1002

AUROC <= 0.5; path length = 2

Brain Area 1				Brain Area 2			
id	acronym	name	parent	id	acronym	name	parent
1114	VISal4	Anterolateral visual area, layer 4	402	1074	VISal1	Anterolateral visual area, layer 1	402
606	RSPv2	Retrosplenial area, ventral part, layer 2	886	622	RSPv6b	Retrosplenial area, ventral part, layer 6b	886
472	MEApd-a	Medial amygdalar nucleus, posterodorsal part, sublayer a	426	480	MEApd-b	Medial amygdalar nucleus, posterodorsal part, sublayer b	426
1072	MGd	Medial geniculate complex, dorsal part	475	1088	MGm	Medial geniculate complex, medial part	475
1088	MGm	Medial geniculate complex, medial part	475	1079	MGv	Medial geniculate complex, ventral part	475
980	PMd	Dorsal premammillary nucleus	467	1004	PMv	Ventral premammillary nucleus	467
559	CEAm	Central amygdalar nucleus, medial part	536	544	CEAc	Central amygdalar nucleus, capsular part	536
281	VISam1	Anteromedial visual area, layer 1	394	1066	VISam2/3	Anteromedial visual area, layer 2/3	394
1042	TTd2	Taenia tecta, dorsal part, layer 2	597	1050	TTd3	Taenia tecta, dorsal part, layer 3	597
148	GU4	Gustatory areas, layer 4	1057	187	GU5	Gustatory areas, layer 5	1057
148	GU4	Gustatory areas, layer 4	1057	662	GU6b	Gustatory areas, layer 6b	1057
783	Ald6a	Agranular insular area, dorsal part, layer 6a	104	1101	Ald5	Agranular insular area, dorsal part, layer 5	104
381	SNr	Substantia nigra, reticular part	323	616	CUN	Cuneiform nucleus	323
74	VISl6a	Lateral visual area, layer 6a	409	973	VISl2/3	Lateral visual area, layer 2/3	409
74	VISl6a	Lateral visual area, layer 6a	409	421	VISl1	Lateral visual area, layer 1	409
416	PAA2	Piriform-amygdalar area, pyramidal layer	788	424	PAA3	Piriform-amygdalar area, polymorph layer	788
868	PBlD	Parabrachial nucleus, lateral division, dorsal lateral part	881	891	PBlv	Parabrachial nucleus, lateral division, ventral lateral part	881
1106	VISC2/3	Visceral area, layer 2/3	677	897	VISC1	Visceral area, layer 1	677
194	LHA	Lateral hypothalamic area	290	173	RCH	Retrochiasmatic area	290
194	LHA	Lateral hypothalamic area	290	226	LPO	Lateral preoptic area	290
194	LHA	Lateral hypothalamic area	290	364	PSTN	Parasubthalamic nucleus	290
368	PERl6b	Perirhinal area, layer 6b	922	692	PERl5	Perirhinal area, layer 5	922
368	PERl6b	Perirhinal area, layer 6b	922	888	PERl2/3	Perirhinal area, layer 2/3	922
1102	SSp-m6a	Primary somatosensory area, mouth, layer 6a	345	2	SSp-m6b	Primary somatosensory area, mouth, layer 6b	345
969	ORBvl1	Orbital area, ventrolateral part, layer 1	746	608	ORBvl6a	Orbital area, ventrolateral part, layer 6a	746
401	VISam4	Anteromedial visual area, layer 4	394	1046	VISam6a	Anteromedial visual area, layer 6a	394
401	VISam4	Anteromedial visual area, layer 4	394	441	VISam6b	Anteromedial visual area, layer 6b	394
540	PERl1	Perirhinal area, layer 1	922	692	PERl5	Perirhinal area, layer 5	922
540	PERl1	Perirhinal area, layer 1	922	888	PERl2/3	Perirhinal area, layer 2/3	922
189	RH	Rhomboid nucleus	51	599	CM	Central medial nucleus of the thalamus	51
880	DTN	Dorsal tegmental nucleus	987	898	PCG	Pontine central gray	987
1066	VISam2/3	Anteromedial visual area, layer 2/3	394	441	VISam6b	Anteromedial visual area, layer 6b	394
255	AV	Anteroventral nucleus of thalamus	239	1113	IAD	Interanterodorsal nucleus of the thalamus	239
965	RSPagl2/3	Retrosplenial area, lateral agranular part, layer 2/3	894	774	RSPagl5	Retrosplenial area, lateral agranular part, layer 5	894
167	AONd	Anterior olfactory nucleus, dorsal part	159	160	AON1	Anterior olfactory nucleus, layer 1	159
167	AONd	Anterior olfactory nucleus, dorsal part	159	183	AONI	Anterior olfactory nucleus, lateral part	159
308	PTLp6a	Posterior parietal association areas, layer 6a	22	340	PTLp6b	Posterior parietal association areas, layer 6b	22
272	AVPV	Anteroventral periventricular nucleus	141	286	SCH	Suprachiasmatic nucleus	141
272	AVPV	Anteroventral periventricular nucleus	141	523	MPO	Medial preoptic area	141
1081	ILA6b	Infralimbic area, layer 6b	44	707	ILA1	Infralimbic area, layer 1	44
1081	ILA6b	Infralimbic area, layer 6b	44	827	ILAS	Infralimbic area, layer 5	44
1081	ILA6b	Infralimbic area, layer 6b	44	556	ILA2/3	Infralimbic area, layer 2/3	44
263	AVP	Anteroventral preoptic nucleus	141	126	PVP	Periventricular hypothalamic nucleus, posterior part	141
1093	PRNc	Pontine reticular nucleus, caudal part	987	534	SUT	Supratrigeminal nucleus	987
240	COApm1	Cortical amygdalar area, posterior part, medial zone, layer 1	663	248	COApm2	Cortical amygdalar area, posterior part, medial zone, layer 2	663
687	RSPv5	Retrosplenial area, ventral part, layer 5	886	622	RSPv6b	Retrosplenial area, ventral part, layer 6b	886
139	ENTl5	Entorhinal area, lateral part, layer 5	918	92	ENTl4	Entorhinal area, lateral part, layer 4	918
1030	SSp-ll1	Primary somatosensory area, lower limb, layer 1	337	113	SSp-ll2/3	Primary somatosensory area, lower limb, layer 2/3	337
574	TRN	Tegmental reticular nucleus	987	534	SUT	Supratrigeminal nucleus	987
837	SUBd-sr	Subiculum, dorsal part, stratum radiatum	509	845	SUBd-sp	Subiculum, dorsal part, pyramidal layer	509
935	ACAd1	Anterior cingulate area, dorsal part, layer 1	39	211	ACAd2/3	Anterior cingulate area, dorsal part, layer 2/3	39
897	VISC1	Visceral area, layer 1	677	857	VISC6a	Visceral area, layer 6a	677

575 CL	Central lateral nucleus of the thalamus	51	599 CM	Central medial nucleus of the thalamus	51
1074 VISal1	Anterolateral visual area, layer 1	402	905 VISal2/3	Anterolateral visual area, layer 2/3	402
1074 VISal1	Anterolateral visual area, layer 1	402	233 VISal5	Anterolateral visual area, layer 5	402
544 CEAc	Central amygdalar nucleus, capsular part	536	551 CEAl	Central amygdalar nucleus, lateral part	536
431 CA2slm	Field CA2, stratum lacunosum-moleculare	423	454 CA2sr	Field CA2, stratum radiatum	423
303 BLAA	Basolateral amygdalar nucleus, anterior part	295	451 BLAV	Basolateral amygdalar nucleus, ventral part	295
266 LSV	Lateral septal nucleus, ventral part	242	258 LSR	Lateral septal nucleus, rostral (rostroventral) part	242
527 AUDd1	Dorsal auditory area, layer 1	1011	600 AUDd2/3	Dorsal auditory area, layer 2/3	1011
646 DP5	Dorsal peduncular area, layer 5	814	496 DP1	Dorsal peduncular area, layer 1	814
118 PVi	Periventricular hypothalamic nucleus, intermediate part	157	223 ARH	Arcuate hypothalamic nucleus	157
616 CUN	Cuneiform nucleus	323	214 RN	Red nucleus	323
772 ACAv5	Anterior cingulate area, ventral part, layer 5	48	810 ACAv6a	Anterior cingulate area, ventral part, 6a	48
269 VISpl2/3	Posterolateral visual area, layer 2/3	425	377 VISpl6a	Posterolateral visual area, layer 6a	425
269 VISpl2/3	Posterolateral visual area, layer 2/3	425	902 VISpl5	Posterolateral visual area, layer 5	425
486 CA3so	Field CA3, stratum oriens	463	471 CA3slm	Field CA3, stratum lacunosum-moleculare	463
1075 TTV2	Taenia tecta, ventral part, layer 2	605	1082 TTV3	Taenia tecta, ventral part, layer 3	605
501 VISpm4	posteromedial visual area, layer 4	533	257 VISpm6a	posteromedial visual area, layer 6a	533
92 ENT14	Entorhinal area, lateral part, layer 4	918	999 ENT12/3	Entorhinal area, lateral part, layer 2/3	918
197 MDc	Mediodorsal nucleus of the thalamus, central part	362	636 MDm	Mediodorsal nucleus of the thalamus, medial part	362
1045 ECT6b	Ectorhinal area/Layer 6b	895	977 ECT6a	Ectorhinal area/Layer 6a	895
28 ENT16a	Entorhinal area, lateral part, layer 6a	918	60 ENT16b	Entorhinal area, lateral part, layer 6b	918
243 AUDd6b	Dorsal auditory area, layer 6b	1011	156 AUDd6a	Dorsal auditory area, layer 6a	1011
52 ENT13	Entorhinal area, lateral part, layer 3	918	999 ENT12/3	Entorhinal area, lateral part, layer 2/3	918
1142 TR3	Postpiriform transition area, layers 3	566	1141 TR2	Postpiriform transition area, layers 2	566
10694 PAR2	Parasubiculum, layer 2	843	10695 PAR3	Parasubiculum, layer 3	843
712 ENTm4	Entorhinal area, medial part, dorsal zone, layer 4	926	664 ENTm3	Entorhinal area, medial part, dorsal zone, layer 3	926
712 ENTm4	Entorhinal area, medial part, dorsal zone, layer 4	926	727 ENTm5	Entorhinal area, medial part, dorsal zone, layer 5	926
565 VISpm5	posteromedial visual area, layer 5	533	257 VISpm6a	posteromedial visual area, layer 6a	533
883 PBIi	Parabrachial nucleus, lateral division, superior lateral part	881	891 PBIv	Parabrachial nucleus, lateral division, ventral lateral part	881
1026 SSp-ul6b	Primary somatosensory area, upper limb, layer 6b	369	945 SSp-ul6a	Primary somatosensory area, upper limb, layer 6a	369
471 CA3slm	Field CA3, stratum lacunosum-moleculare	463	479 CA3slu	Field CA3, stratum lucidum	463
471 CA3slm	Field CA3, stratum lacunosum-moleculare	463	495 CA3sp	Field CA3, pyramidal layer	463
872 DR	Dorsal nucleus raphe	165	100 IPN	Interpeduncular nucleus	165
460 MEV	Midbrain trigeminal nucleus	339	580 NB	Nucleus of the brachium of the inferior colliculus	339
162 LDT	Laterodorsal tegmental nucleus	1117	358 SLD	Sublaterodorsal nucleus	1117
757 VTN	Ventral tegmental nucleus	323	246 RR	Midbrain reticular nucleus, retrorubral area	323
523 MPO	Medial preoptic area	141	347 SBPV	Subparaventricular zone	141
523 MPO	Medial preoptic area	141	126 PVP	Periventricular hypothalamic nucleus, posterior part	141
1010 VISC4	Visceral area, layer 4	677	1058 VISC5	Visceral area, layer 5	677
1105 IA	Intercalated amygdalar nucleus	278	23 AAA	Anterior amygdalar area	278
149 PVT	Paraventricular nucleus of the thalamus	571	15 PT	Parataenial nucleus	571
604 NI	Nucleus incertus	1117	238 RPO	Nucleus raphe pontis	1117
564 MS	Medial septal nucleus	904	596 NDB	Diagonal band nucleus	904
860 PBIc	Parabrachial nucleus, lateral division, central lateral part	881	875 PBIe	Parabrachial nucleus, lateral division, external lateral part	881
860 PBIc	Parabrachial nucleus, lateral division, central lateral part	881	891 PBIv	Parabrachial nucleus, lateral division, ventral lateral part	881
800 Alv5	Agranular insular area, ventral part, layer 5	119	704 Alv1	Agranular insular area, ventral part, layer 1	119
907 PCN	Paracentral nucleus	51	599 CM	Central medial nucleus of the thalamus	51
335 PER16a	Perirhinal area, layer 6a	922	888 PER12/3	Perirhinal area, layer 2/3	922
84 PL6a	Prelimbic area, layer 6a	972	363 PL5	Prelimbic area, layer 5	972
724 AHNp	Anterior hypothalamic nucleus, posterior part	88	708 AHNc	Anterior hypothalamic nucleus, central part	88
347 SBPV	Subparaventricular zone	141	126 PVP	Periventricular hypothalamic nucleus, posterior part	141
591 CLI	Central linear nucleus raphe	165	100 IPN	Interpeduncular nucleus	165
973 VISI2/3	Lateral visual area, layer 2/3	409	421 VISI1	Lateral visual area, layer 1	409
810 ACAv6a	Anterior cingulate area, ventral part, 6a	48	588 ACAv1	Anterior cingulate area, ventral part, layer 1	48
700 AHNa	Anterior hypothalamic nucleus, anterior part	88	708 AHNc	Anterior hypothalamic nucleus, central part	88
232 COApl3	Cortical amygdalar area, posterior part, lateral zone, layer 3	655	224 COApl2	Cortical amygdalar area, posterior part, lateral zone, layer 2	655
377 VISpl6a	Posterolateral visual area, layer 6a	425	902 VISpl5	Posterolateral visual area, layer 5	425
1125 ORBvl5	Orbital area, ventrolateral part, layer 5	746	288 ORBvl2/3	Orbital area, ventrolateral part, layer 2/3	746
845 SUBd-sp	Subiculum, dorsal part, pyramidal layer	509	829 SUBd-m	Subiculum, dorsal part, molecular layer	509
1015 ACAd5	Anterior cingulate area, dorsal part, layer 5	39	919 ACAd6a	Anterior cingulate area, dorsal part, layer 6a	39
216 COApl1	Cortical amygdalar area, posterior part, lateral zone, layer 1	655	224 COApl2	Cortical amygdalar area, posterior part, lateral zone, layer 2	655
613 VISI5	Lateral visual area, layer 5	409	421 VISI1	Lateral visual area, layer 1	409
10693 PAR1	Parasubiculum, layer 1	843	10695 PAR3	Parasubiculum, layer 3	843
56 ACB	Nucleus accumbens	493	998 FS	Fundus of striatum	493
10701 PRE3	Presubiculum, layer 3	1084	10700 PRE2	Presubiculum, layer 2	1084
664 ENTm3	Entorhinal area, medial part, dorsal zone, layer 3	926	727 ENTm5	Entorhinal area, medial part, dorsal zone, layer 5	926
358 SLD	Sublaterodorsal nucleus	1117	238 RPO	Nucleus raphe pontis	1117
268 NLOT2	Nucleus of the lateral olfactory tract, pyramidal layer	619	1139 NLOT3	Nucleus of the lateral olfactory tract, layer 3	619
268 NLOT2	Nucleus of the lateral olfactory tract, pyramidal layer	619	260 NLOT1	Nucleus of the lateral olfactory tract, molecular layer	619
479 CA3slu	Field CA3, stratum lucidum	463	495 CA3sp	Field CA3, pyramidal layer	463
511 SCig-c	Superior colliculus, motor related, intermediate gray layer, su	10	494 SCig-a	Superior colliculus, motor related, intermediate gray layer, sul	10
310 SF	Septofimbrial nucleus	275	333 SH	Septohippocampal nucleus	275
484 ORBm1	Orbital area, medial part, layer 1	731	620 ORBm5	Orbital area, medial part, layer 5	731
638 GU6a	Gustatory areas, layer 6a	1057	662 GU6b	Gustatory areas, layer 6b	1057
15 PT	Parataenial nucleus	571	181 RE	Nucleus of reunions	571
478 SSp-ll6a	Primary somatosensory area, lower limb, layer 6a	337	510 SSp-ll6b	Primary somatosensory area, lower limb, layer 6b	337
598 AUDv6b	Ventral auditory area, layer 6b	1018	1023 AUDv5	Ventral auditory area, layer 5	1018

137	CSI	Superior central nucleus raphe, lateral part	679	130	CSm	Superior central nucleus raphe, medial part	679
727	ENTm5	Entorhinal area, medial part, dorsal zone, layer 5	926	743	ENTm6	Entorhinal area, medial part, dorsal zone, layer 6	926
450	SSp-ul1	Primary somatosensory area, upper limb, layer 1	369	854	SSp-ul2/3	Primary somatosensory area, upper limb, layer 2/3	369
1139	NLOT3	Nucleus of the lateral olfactory tract, layer 3	619	260	NLOT1	Nucleus of the lateral olfactory tract, molecular layer	619
440	ORBI6a	Orbital area, lateral part, layer 6a	723	630	ORBI5	Orbital area, lateral part, layer 5	723

Supplementary Table 3

ABA to ST LASSO, alpha=0.1

auroc file = "ABAtoST_Tall_f1_Op1_051420.csv"

AUROC >= 0.95; path length = 2							
Brain Area 1				Brain Area 2			
id	acronym	name	parent	id	acronym	name	parent
1005	AUDp6b	Primary auditory area, layer 6b	1002	816	AUDp4	Primary auditory area, layer 4	1002
943	MOp2/3	Primary motor area, Layer 2/3	985	882	MOp6b	Primary motor area, Layer 6b	985
882	MOp6b	Primary motor area, Layer 6b	985	320	MOp1	Primary motor area, Layer 1	985
269	VISpl2/3	Posterolateral visual area, layer 2/3	425	377	VISpl6a	Posterolateral visual area, layer 6a	425
1045	ECT6b	Ectorhinal area/Layer 6b	895	836	ECT1	Ectorhinal area/Layer 1	895

AUROC <=0.5; path length = 2							
Brain Area 1				Brain Area 2			
id	acronym	name	parent	id	acronym	name	parent
657	SSp-m2/3	Primary somatosensory area, mouth, layer 2/3	345	950	SSp-m4	Primary somatosensory area, mouth, layer 4	345
1114	VISal4	Anterolateral visual area, layer 4	402	233	VISal5	Anterolateral visual area, layer 5	402
606	RSPv2	Retrosplenial area, ventral part, layer 2	886	622	RSPv6b	Retrosplenial area, ventral part, layer 6b	886
606	RSPv2	Retrosplenial area, ventral part, layer 2	886	430	RSPv2/3	Retrosplenial area, ventral part, layer 2/3	886
472	MEApd-a	Medial amygdalar nucleus, posterodorsal part, sublayer a	426	487	MEApd-c	Medial amygdalar nucleus, posterodorsal part, sublayer c	426
472	MEApd-a	Medial amygdalar nucleus, posterodorsal part, sublayer a	426	480	MEApd-b	Medial amygdalar nucleus, posterodorsal part, sublayer b	426
980	PMd	Dorsal premammillary nucleus	467	946	PH	Posterior hypothalamic nucleus	467
980	PMd	Dorsal premammillary nucleus	467	1004	PMv	Ventral premammillary nucleus	467
296	ACAv2/3	Anterior cingulate area, ventral part, layer 2/3	48	772	ACAv5	Anterior cingulate area, ventral part, layer 5	48
148	GU4	Gustatory areas, layer 4	1057	187	GU5	Gustatory areas, layer 5	1057
148	GU4	Gustatory areas, layer 4	1057	662	GU6b	Gustatory areas, layer 6b	1057
783	Ald6a	Agranular insular area, dorsal part, layer 6a	104	1101	Ald5	Agranular insular area, dorsal part, layer 5	104
381	SNr	Substantia nigra, reticular part	323	616	CUN	Cuneiform nucleus	323
381	SNr	Substantia nigra, reticular part	323	757	VTN	Ventral tegmental nucleus	323
191	AONm	Anterior olfactory nucleus, medial part	159	167	AONd	Anterior olfactory nucleus, dorsal part	159
416	PAA2	Piriform-amygdalar area, pyramidal layer	788	424	PAA3	Piriform-amygdalar area, polymorph layer	788
868	PBlid	Parabrachial nucleus, lateral division, dorsal lateral part	881	860	PBlc	Parabrachial nucleus, lateral division, central lateral part	881
868	PBlid	Parabrachial nucleus, lateral division, dorsal lateral part	881	891	PBlv	Parabrachial nucleus, lateral division, ventral lateral part	881
1106	VISC2/3	Visceral area, layer 2/3	677	1058	VISC5	Visceral area, layer 5	677
628	NOT	Nucleus of the optic tract	1100	634	NPC	Nucleus of the posterior commissure	1100
628	NOT	Nucleus of the optic tract	1100	215	APN	Anterior prepectal nucleus	1100
105	SOCm	Superior olivary complex, medial part	398	122	POR	Superior olivary complex, periolivary region	398
194	LHA	Lateral hypothalamic area	290	364	PSTN	Parasubthalamic nucleus	290
1062	SSp-bfd6b	Primary somatosensory area, barrel field, layer 6b	329	1070	SSp-bfd5	Primary somatosensory area, barrel field, layer 5	329
465	OT2	Olfactory tubercle, pyramidal layer	754	473	OT3	Olfactory tubercle, polymorph layer	754
465	OT2	Olfactory tubercle, pyramidal layer	754	481	isl	Islands of Calleja	754
1102	SSp-m6a	Primary somatosensory area, mouth, layer 6a	345	878	SSp-m1	Primary somatosensory area, mouth, layer 1	345
1102	SSp-m6a	Primary somatosensory area, mouth, layer 6a	345	2	SSp-m6b	Primary somatosensory area, mouth, layer 6b	345
189	RH	Rhomboid nucleus	51	575	CL	Central lateral nucleus of the thalamus	51
694	Aiv2/3	Agranular insular area, ventral part, layer 2/3	119	800	Aiv5	Agranular insular area, ventral part, layer 5	119
344	Alp5	Agranular insular area, posterior part, layer 5	111	314	Alp6a	Agranular insular area, posterior part, layer 6a	111
965	RSPagl2/3	Retrosplenial area, lateral agranular part, layer 2/3	894	774	RSPagl5	Retrosplenial area, lateral agranular part, layer 5	894
272	AVPV	Anteroventral periventricular nucleus	141	523	MPO	Medial preoptic area	141
272	AVPV	Anteroventral periventricular nucleus	141	347	SBPV	Subparaventricular zone	141
263	AVP	Anteroventral preoptic nucleus	141	286	SCH	Suprachiasmatic nucleus	141
263	AVP	Anteroventral preoptic nucleus	141	523	MPO	Medial preoptic area	141
263	AVP	Anteroventral preoptic nucleus	141	126	PVp	Periventricular hypothalamic nucleus, posterior part	141
458	OT1	Olfactory tubercle, molecular layer	754	481	isl	Islands of Calleja	754
687	RSPv5	Retrosplenial area, ventral part, layer 5	886	430	RSPv2/3	Retrosplenial area, ventral part, layer 2/3	886
292	BA	Bed nucleus of the accessory olfactory tract	278	1105	IA	Intercalated amygdalar nucleus	278
635	PTLp4	Posterior parietal association areas, layer 4	22	241	PTLp2/3	Posterior parietal association areas, layer 2/3	22
683	PTLp5	Posterior parietal association areas, layer 5	22	241	PTLp2/3	Posterior parietal association areas, layer 2/3	22
622	RSPv6b	Retrosplenial area, ventral part, layer 6b	886	590	RSPv6a	Retrosplenial area, ventral part, layer 6a	886
622	RSPv6b	Retrosplenial area, ventral part, layer 6b	886	430	RSPv2/3	Retrosplenial area, ventral part, layer 2/3	886
1086	SSp-tr4	Primary somatosensory area, trunk, layer 4	361	461	SSp-tr6b	Primary somatosensory area, trunk, layer 6b	361
305	VISp6b	Primary visual area, layer 6b	385	33	VISp6a	Primary visual area, layer 6a	385
837	SUBd-sr	Subiculum, dorsal part, stratum radiatum	509	845	SUBd-sp	Subiculum, dorsal part, pyramidal layer	509
544	CEAc	Central amygdalar nucleus, capsular part	536	551	CEAl	Central amygdalar nucleus, lateral part	536
411	MEAad	Medial amygdalar nucleus, anterodorsal part	403	418	MEAav	Medial amygdalar nucleus, anteroventral part	403
614	TU	Tuberal nucleus	290	173	RCH	Retrochiasmatic area	290
614	TU	Tuberal nucleus	290	226	LPO	Lateral preoptic area	290
187	GU5	Gustatory areas, layer 5	1057	662	GU6b	Gustatory areas, layer 6b	1057
41	VISpm2/3	posteromedial visual area, layer 2/3	533	565	VISpm5	posteromedial visual area, layer 5	533
303	BLAa	Basolateral amygdalar nucleus, anterior part	295	451	BLAv	Basolateral amygdalar nucleus, ventral part	295
654	SSp-n4	Primary somatosensory area, nose, layer 4	353	838	SSp-n2/3	Primary somatosensory area, nose, layer 2/3	353
266	LSv	Lateral septal nucleus, ventral part	242	258	LSr	Lateral septal nucleus, rostral (rostroventral) part	242
304	PL2/3	Prelimbic area, layer 2/3	972	363	PL5	Prelimbic area, layer 5	972
646	DP5	Dorsal peduncular area, layer 5	814	360	DP2/3	Dorsal peduncular area, layer 2/3	814
412	ORB12/3	Orbital area, lateral part, layer 2/3	723	448	ORB1	Orbital area, lateral part, layer 1	723
616	CUN	Cuneiform nucleus	323	757	VTN	Ventral tegmental nucleus	323
772	ACAv5	Anterior cingulate area, ventral part, layer 5	48	588	ACAv1	Anterior cingulate area, ventral part, layer 1	48
286	SCH	Suprachiasmatic nucleus	141	347	SBPV	Subparaventricular zone	141
486	CA3so	Field CA3, stratum oriens	463	471	CA3sm	Field CA3, stratum lacunosum-moleculare	463
1075	TTv2	Taenia tecta, ventral part, layer 2	605	1082	TTv3	Taenia tecta, ventral part, layer 3	605
692	PERI5	Perirhinal area, layer 5	922	335	PERI6a	Perirhinal area, layer 6a	922
501	VISpm4	posteromedial visual area, layer 4	533	565	VISpm5	posteromedial visual area, layer 5	533
233	VISal5	Anterolateral visual area, layer 5	402	649	VISal6b	Anterolateral visual area, layer 6b	402
233	VISal5	Anterolateral visual area, layer 5	402	601	VISal6a	Anterolateral visual area, layer 6a	402
520	AUDv6a	Ventral auditory area, layer 6a	1018	598	AUDv6b	Ventral auditory area, layer 6b	1018
1045	ECT6b	Ectorhinal area/Layer 6b	895	977	ECT6a	Ectorhinal area/Layer 6a	895
52	ENT13	Entorhinal area, lateral part, layer 3	918	715	ENT12a	Entorhinal area, lateral part, layer 2a	918

473 OT3	Olfactory tubercle, polymorph layer	754	481 isl	Islands of Calleja	754
712 ENTm4	Entorhinal area, medial part, dorsal zone, layer 4	926	664 ENTm3	Entorhinal area, medial part, dorsal zone, layer 3	926
883 PBIs	Parabrachial nucleus, lateral division, superior lateral part	881	891 PBlv	Parabrachial nucleus, lateral division, ventral lateral part	881
471 CA3slm	Field CA3, stratum lacunosum-moleculare	463	495 CA3sp	Field CA3, pyramidal layer	463
670 SSp-tr2/3	Primary somatosensory area, trunk, layer 2/3	361	461 SSp-tr6b	Primary somatosensory area, trunk, layer 6b	361
872 DR	Dorsal nucleus raphe	165	591 CLI	Central linear nucleus raphe	165
460 MEV	Midbrain trigeminal nucleus	339	580 NB	Nucleus of the brachium of the inferior colliculus	339
537 BSTal	Bed nuclei of the stria terminalis, anterior division, anterolateral area	359	498 BSTam	Bed nuclei of the stria terminalis, anterior division, anteromedial area	359
1094 SSp-II4	Primary somatosensory area, lower limb, layer 4	337	510 SSp-II6b	Primary somatosensory area, lower limb, layer 6b	337
162 LDT	Laterodorsal tegmental nucleus	1117	358 SLD	Sublaterodorsal nucleus	1117
162 LDT	Laterodorsal tegmental nucleus	1117	238 RPO	Nucleus raphe pontis	1117
757 VTN	Ventral tegmental nucleus	323	749 VTA	Ventral tegmental area	323
757 VTN	Ventral tegmental nucleus	323	246 RR	Midbrain reticular nucleus, retrorubral area	323
757 VTN	Ventral tegmental nucleus	323	214 RN	Red nucleus	323
889 SSp-n6a	Primary somatosensory area, nose, layer 6a	353	702 SSp-n5	Primary somatosensory area, nose, layer 5	353
149 PVT	Paraventricular nucleus of the thalamus	571	15 PT	Parataenial nucleus	571
604 NI	Nucleus incertus	1117	358 SLD	Sublaterodorsal nucleus	1117
307 MARN	Magnocellular reticular nucleus	370	661 VII	Facial motor nucleus	370
907 PCN	Paracentral nucleus	51	599 CM	Central medial nucleus of the thalamus	51
649 VISal6b	Anterolateral visual area, layer 6b	402	601 VISal6a	Anterolateral visual area, layer 6a	402
724 AHNp	Anterior hypothalamic nucleus, posterior part	88	708 AHNc	Anterior hypothalamic nucleus, central part	88
591 CLI	Central linear nucleus raphe	165	100 IPN	Interpeduncular nucleus	165
487 MEApd-c	Medial amygdalar nucleus, posterodorsal part, sublayer c	426	480 MEApd-b	Medial amygdalar nucleus, posterodorsal part, sublayer b	426
232 COAp13	Cortical amygdalar area, posterior part, lateral zone, layer 3	655	224 COAp12	Cortical amygdalar area, posterior part, lateral zone, layer 2	655
377 VISpl6a	Posterolateral visual area, layer 6a	425	902 VISpl5	Posterolateral visual area, layer 5	425
454 CA2sr	Field CA2, stratum radiatum	423	446 CA2sp	Field CA2, pyramidal layer	423
1113 IAD	Interanterodorsal nucleus of the thalamus	239	155 LD	Lateral dorsal nucleus of thalamus	239
503 SCig-b	Superior colliculus, motor related, intermediate gray layer, sublayer b	10	511 SCig-c	Superior colliculus, motor related, intermediate gray layer, sublayer c	10
634 NPC	Nucleus of the posterior commissure	1100	215 APN	Anterior pretectal nucleus	1100
613 VISI5	Lateral visual area, layer 5	409	421 VISI1	Lateral visual area, layer 1	409
56 ACB	Nucleus accumbens	493	998 FS	Fundus of striatum	493
578 BSTpr	Bed nuclei of the stria terminalis, posterior division, principal nucleus	367	585 BSTif	Bed nuclei of the stria terminalis, posterior division, interfascicular nucleus	367
676 DMHp	Dorsomedial nucleus of the hypothalamus, posterior part	830	668 DMHa	Dorsomedial nucleus of the hypothalamus, anterior part	830
360 DP2/3	Dorsal peduncular area, layer 2/3	814	496 DP1	Dorsal peduncular area, layer 1	814
479 CA3slu	Field CA3, stratum lucidum	463	495 CA3sp	Field CA3, pyramidal layer	463
511 SCig-c	Superior colliculus, motor related, intermediate gray layer, sublayer c	10	494 SCig-a	Superior colliculus, motor related, intermediate gray layer, sublayer a	10
778 VISp5	Primary visual area, layer 5	385	721 VISp4	Primary visual area, layer 4	385
310 SF	Septofimbrial nucleus	275	333 SH	Septohippocampal nucleus	275
638 GU6a	Gustatory areas, layer 6a	1057	662 GU6b	Gustatory areas, layer 6b	1057
764 ENT12b	Entorhinal area, lateral part, layer 2b	918	715 ENT12a	Entorhinal area, lateral part, layer 2a	918
1046 VISam6a	Anteromedial visual area, layer 6a	394	441 VISam6b	Anteromedial visual area, layer 6b	394
668 DMHa	Dorsomedial nucleus of the hypothalamus, anterior part	830	684 DMHv	Dorsomedial nucleus of the hypothalamus, ventral part	830
1127 TEa2/3	Temporal association areas, layer 2/3	541	234 TEa4	Temporal association areas, layer 4	541
875 PBlc	Parabrachial nucleus, lateral division, external lateral part	881	891 PBlv	Parabrachial nucleus, lateral division, ventral lateral part	881
1096 AMd	Anteromedial nucleus, dorsal part	127	1104 AMv	Anteromedial nucleus, ventral part	127
440 ORBl6a	Orbital area, lateral part, layer 6a	723	630 ORB15	Orbital area, lateral part, layer 5	723






Appendix D

Nature Neuroscience Technical Report: Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections

This appendix contains the full text of the Nature Neuroscience Technical Report titled "Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections," which was authored jointly by Yu-Chi Sun, Xiaoyin Chen, Stephan Fischer, Shaina Lu, Huiqing Zhan, Jesse Gillis, and Anthony M. Zador. I performed the comparison of BARseq data with the Allen Brain Atlas in Figure 2f and Extended Data Figure 3.



Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections

Yu-Chi Sun^{1,2}, Xiaoyin Chen^{1,2}  , Stephan Fischer¹, Shaina Lu¹ , Huiqing Zhan¹, Jesse Gillis¹ and Anthony M. Zador¹  

Functional circuits consist of neurons with diverse axonal projections and gene expression. Understanding the molecular signature of projections requires high-throughput interrogation of both gene expression and projections to multiple targets in the same cells at cellular resolution, which is difficult to achieve using current technology. Here, we introduce BARseq2, a technique that simultaneously maps projections and detects multiplexed gene expression by in situ sequencing. We determined the expression of cadherins and cell-type markers in 29,933 cells and the projections of 3,164 cells in both the mouse motor cortex and auditory cortex. Associating gene expression and projections in 1,349 neurons revealed shared cadherin signatures of homologous projections across the two cortical areas. These cadherins were enriched across multiple branches of the transcriptomic taxonomy. By correlating multigene expression and projections to many targets in single neurons with high throughput, BARseq2 provides a potential path to uncovering the molecular logic underlying neuronal circuits.

Neural circuits are composed of neurons diverse in many properties, such as morphology^{1,2}, gene expression^{3,4} and projections^{5,6}. Although recent technological advances have made it possible to characterize the diversity in individual neuronal properties, associating multiple properties in single neurons with high throughput remains difficult to achieve. Investigating the relationship between multiple neuronal properties is essential for understanding the complex organization of neural circuits.

Of particular interest is the relationship between endogenous gene expression and long-range projections in the cortex. Cortical neurons have diverse patterns of long-range projections^{5,6} and diverse patterns of gene expression^{3,4}. The full diversity of neuronal projection patterns can often only be appreciated by assessing multiple projection targets simultaneously (Fig. 1a)^{2,6}. For example, Han et al.⁵ showed that neurons in mouse visual area V1 that project to area PM tend not to project to area AL and vice versa, a projection ‘motif’ that involves the relative probability that a single neuron projects to two targets and hence could not have been discovered by assessing projection targets one at a time. Gene expression patterns are also complex, and although the diversity in gene expression can be described by clustering neurons into transcriptomic types, these transcriptomic types have limited power in explaining the diversity of cortical projections beyond the major classes of projection neurons^{3,6–8}. Moreover, because the determination of a transcriptomic type relies on the expression of only a subset of genes, the inability of transcriptomic type to predict projection patterns raises the possibility that the expression of other genes—potentially in gene coexpression motifs—might be better correlated with projection patterns. Although transcriptomic methods can be combined with retrograde labeling^{3,9}, retrograde labeling is limited to one or at most a few brain areas at a time. Resolving the relationship between gene expression and projection patterns in the adult cortex thus requires high-throughput techniques that allow simultaneous multiplexed

gene detection with projection mapping to multiple target areas at single-neuron resolution, which remains difficult to achieve.

To achieve high-throughput mapping of projections to many brain areas, we recently introduced barcoded anatomy resolved by sequencing (BARseq), a projection mapping technique based on in situ sequencing of RNA barcodes⁶. In BARseq, each neuron is labeled with a unique virally encoded RNA barcode that is replicated in the somas and transported to the axon terminals. The barcodes at the axon terminals located at various target areas are sequenced and matched to somatic barcodes, which are sequenced in situ, to determine the projection patterns of each labeled neuron. Because BARseq preserves the location of somata with high spatial resolution, in principle it provides a platform to combine projection mapping with other neuronal properties also interrogated in situ, including gene expression. We have previously shown⁶ that BARseq can be combined with fluorescence in situ hybridization (FISH) and *Cre* labeling to uncover projections across neuronal subtypes defined by gene expression. However, these approaches can only interrogate one or a few genes at a time, which would be insufficient for unraveling the complex relationship between the expression of many genes to diverse cortical projections (Fig. 1a).

Here we aim to develop a technique to simultaneously map projections to multiple brain areas and detect the expression of dozens of genes in hundreds to thousands of neurons from a cortical area with high throughput, high spatial resolution and cellular resolution. To achieve this goal, we combine the high-throughput and multiplexed projection mapping capability of BARseq with state-of-the-art spatial transcriptomic techniques with high imaging throughput and multiplexing capacity^{10,11}. This second-generation BARseq (BARseq2) greatly improves the ability to correlate the expression of many genes to projections to many targets in the same neurons. As a proof of principle, we first demonstrate multiplexed gene detection using BARseq2 by mapping the spatial pattern of up

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. ²These authors contributed equally: Yu-Chi Sun, Xiaoyin Chen. ✉e-mail: xichen@cshl.edu; zador@cshl.edu

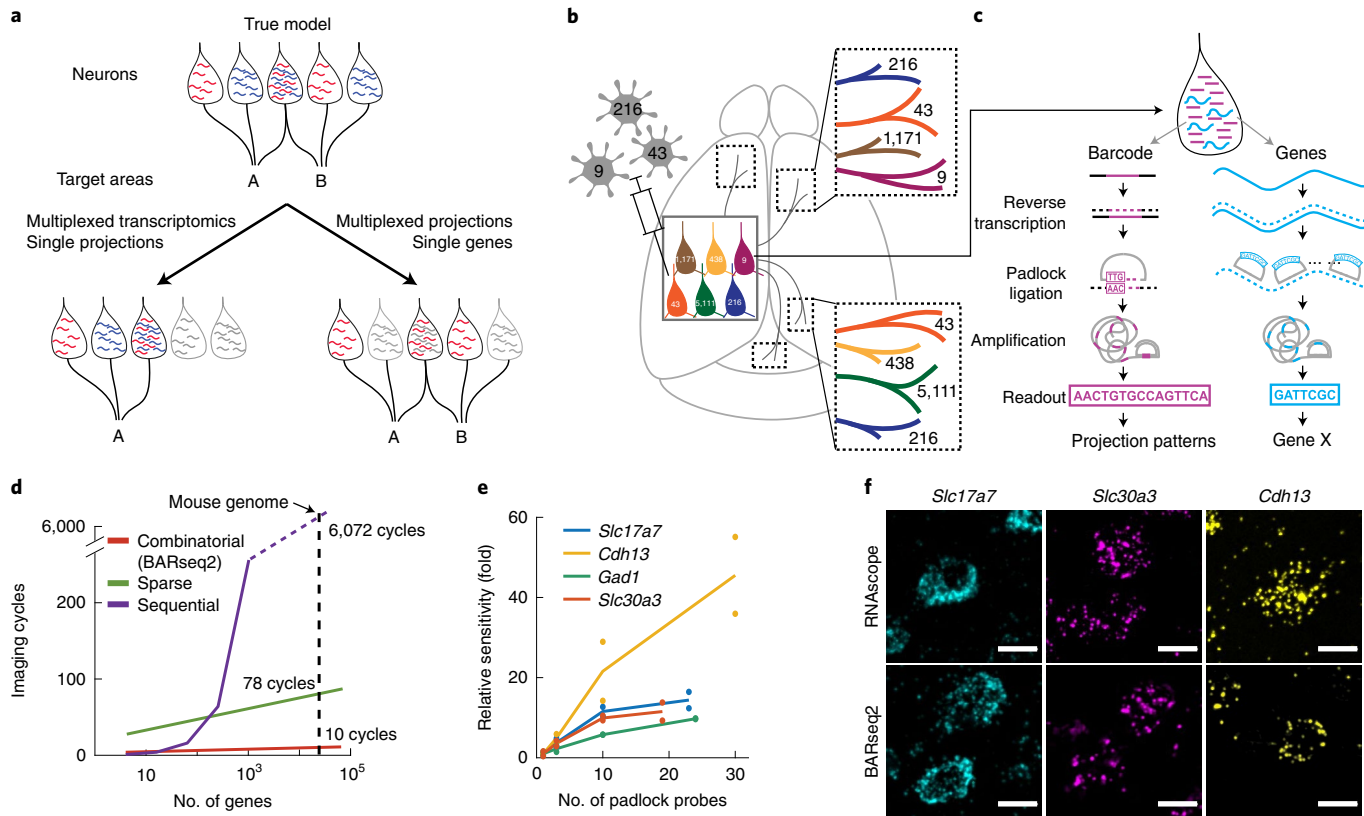


Fig. 1 | In situ sequencing of endogenous mRNAs using BARseq2. **a**, Cartoon of an example model in which the relationship between projections and gene expression can only be correctly inferred by multiplexed interrogation of both projections and gene expression. In this model, neurons that express both genes project to both targets A and B, whereas neurons that express only one of the two genes project randomly to either A or B, but not both. Methods that combine multiplexed single-neuron gene expression with data about only a single projection target will conclude that all three gene expression patterns project to target A, thus failing to detect the underlying ‘true’ relationship between gene expression and projections. Similarly, methods that combine multiplexed single-neuron projections with data about only a single gene will also fail to detect any relationship between gene expression and projections. **b,c**, BARseq2 correlates projections and gene expression at cellular resolution (**b**). In BARseq2, neurons are barcoded with random RNA sequences to allow projection mapping, and genes are also sequenced in the same barcoded neurons. RNA barcodes and genes are amplified and read out using different strategies (**c**). **d**, Theoretical imaging cycles using combinatorial coding (BARseq2), four-channel sequential coding or four-channel sparse coding as used by Eng et al.⁵⁰. Imaging cycles assumed three additional cycles for BARseq2, one additional round for sparse coding, and no extra cycle for sequential coding for error correction. **e**, Mean and individual data points of the relative sensitivity of BARseq2 in detecting the indicated genes using different numbers of padlock probes per gene. The sensitivity is normalized to that using one probe per gene. $n = 2$ slices for each gene. **f**, Representative images of BARseq2 detection of the indicated genes using the maximum number of probes shown in **e** compared to RNAscope. Scale bars, 10 μm .

to 65 cadherins and cell-type markers in 29,933 cells. We then correlate the expression of 20 cadherins to projections to 35 target areas in 1,349 neurons in mouse motor and auditory cortex. Our study reveals new sets of cadherins that correlate with homologous projections in both cortical areas. BARseq2 thus bridges transcriptomic signatures obtained through spatial transcriptional profiling with sequencing-based projection mapping to illuminate the molecular logic of long-range projections.

Results

To investigate how cadherin expression relates to diverse projections, we developed BARseq2 to combine high-throughput projection mapping with multiplexed detection of gene expression using in situ sequencing (Fig. 1b,c). BARseq2 is based on BARseq (Fig. 1c), which achieves high-throughput projection mapping by in situ sequencing of RNA barcodes⁶. Projection patterns observed using BARseq are consistent with those obtained using conventional neuroanatomical techniques in multiple circuits^{2,5}, but it can achieve throughput that is at least two to three orders of magnitude higher than the state-of-the-art single-cell tracing techniques². Possible technical concerns, including distinguishing fibers of passage from

axonal termini, sensitivity, double labeling of neurons and degenerate barcodes, have previously been addressed^{2,6,12,13} and will not be discussed in detail again here. Combining barcoded single-cell projection mapping with in situ detection of endogenous mRNAs exploits the unique advantage of BARseq in throughput to efficiently interrogate both neuronal gene expression and long-range projections simultaneously.

To detect gene expression using BARseq2, we used a non-gap-filled padlock probe-based approach to amplify target endogenous mRNAs^{10,11} (Fig. 1c). The elimination of gap filling, necessary for reading out extremely diverse sequences of barcodes, increases the sensitivity for endogenous gene detection. In this approach, the identity of the target is read out by sequencing a gene-identification index (GII) using Illumina sequencing chemistry in situ. Because the GII is a nucleotide barcode sequence that uniquely encodes the identity of a given gene, the multiplexing capacity increases exponentially as 4^N , where N is the number of sequencing cycles. This combinatorial coding by sequencing readout thereby allows simultaneous detection of a large number of genes using only a few cycles of imaging (Fig. 1d). Although sequencing readout offers many advantages, BARseq2 is also compatible with hybridization-based

readout when necessary. The combination of non-gap-filling in situ sequencing of endogenous genes and the gap-filling approach for sequencing barcodes allows many genes to be detected simultaneously with projections using BARseq2.

We first demonstrate that, by optimizing targeted in situ sequencing, BARseq2 could achieve sufficient sensitivity for detection of endogenous mRNAs. We next combined in situ sequencing of endogenous mRNAs with in situ sequencing of RNA barcodes to associate the expression of cadherins with projection patterns at cellular resolution. We then validated BARseq2 by demonstrating that it could be used to recapitulate projection patterns specific to transcriptomic neuronal subtypes and to identify cadherins that were differentially expressed across major projection classes. Finally, we identified a set of cadherins shared between the mouse auditory cortex and motor cortex that correlate with homologous projections of intratelencephalic (IT) neurons in both cortical areas.

BARseq2 robustly detects endogenous mRNAs. To adequately detect genes using BARseq2, we sought to improve the detection sensitivity. In most in situ hybridization (ISH) methods, high sensitivity is achieved by using many probes for each target mRNA^{14,15}. We reasoned that increasing the number of padlock probes for each gene might similarly improve the sensitivity of BARseq2. Indeed, we observed that tiling the whole gene with additional probes resulted in as much as a 46-fold increase in sensitivity compared to using a single probe (Fig. 1e and Methods). Combined with other technical optimizations (Extended Data Fig. 1a,b), we increased the sensitivity of BARseq2 to 60% of RNAscope, a sensitive and commercially available FISH method (Fig. 1f, Extended Data Fig. 1c,d and Methods). We further optimized in situ sequencing to robustly read out GIs of single rlonies over many sequencing cycles (Extended Data Fig. 1e–j and Methods). These optimizations allowed BARseq2 to achieve sufficiently sensitive, fast and robust detection of mRNAs.

BARseq2 allows multiplexed detection of mRNAs in situ. To assess multiplexed detection of cadherins in situ using BARseq2, we examined the expression of 20 cadherins, along with either 3 (in auditory cortex) or 45 (in motor cortex) cell-type markers (Fig. 2a–c). We chose to focus on the cadherins because of their known roles in cortical development, including projection specification^{16,17}, and their differential expression among cardinal cell types defined by multiple properties¹⁸. These cadherins included most classical cadherins and nonclustered protocadherins expressed in auditory cortex and motor cortex. We successfully resolved and decoded 419,724 rlonies from two slices of mouse auditory cortex (1.7 mm² × 10 μm per slice) and 1,445,648 rlonies from four slices of primary motor cortex (2.8 mm² × 10 μm per slice). We recovered 20 rlonies in auditory cortex and 115 rlonies in motor cortex matching two GIs that were not used in the experiment,

corresponding to an estimated error rate of 0.1% and 0.2%, respectively, for rlonie decoding.

Consistent with previous reports^{19,20}, many cadherins were enriched in specific layers and sublayers in the cortex (Fig. 2d). Interestingly, although most cadherins had similar laminar expression in both auditory cortex and motor cortex, some cadherins were differentially expressed across the two areas. For example, *Cdh9* and *Cdh13* were enriched in L2/3 in auditory cortex, but not in motor cortex (Fig. 2d and Extended Data Fig. 2). The laminar positions of peak cadherin expression were consistent with those obtained by other methods, including RNAscope (Fig. 2e) and the Allen Brain Atlas (ABA) database of ISH²¹ (Fig. 2f, Extended Data Fig. 3 and Methods). Thus, BARseq2 accurately resolved the laminar expression patterns of cadherins.

We then characterized gene expression obtained by BARseq2 at single-cell resolution (Methods). We assigned 228,371 rlonies to 3,377 excitatory or inhibitory neurons (67.6 ± 28.8 (mean ± s.d.) rlonies per neuron) in auditory cortex, and 752,687 rlonies to 11,492 excitatory or inhibitory neurons (65.5 ± 26.0 (mean ± s.d.) rlonies per neuron) in motor cortex. Most cadherins showed slight differences in single-cell expression levels in these two cortical areas (Extended Data Fig. 4). In auditory cortex, the total read counts per cell was higher in BARseq2 than in single-cell RNA sequencing (scRNA-seq) using 10x Genomics v3 (Fig. 2g; median read count was 64 for BARseq2 ($n = 3,337$ cells) compared to 57 for scRNA-seq ($n = 640$ cells); $P = 5.3 \times 10^{-5}$, rank-sum test). Thus, even using a limited number of probes, BARseq2 achieved sensitivity at least equal to scRNA-seq using 10x v3. For experiments requiring better quantification of genes present at low expression, the sensitivity could potentially be further improved by using more probes.

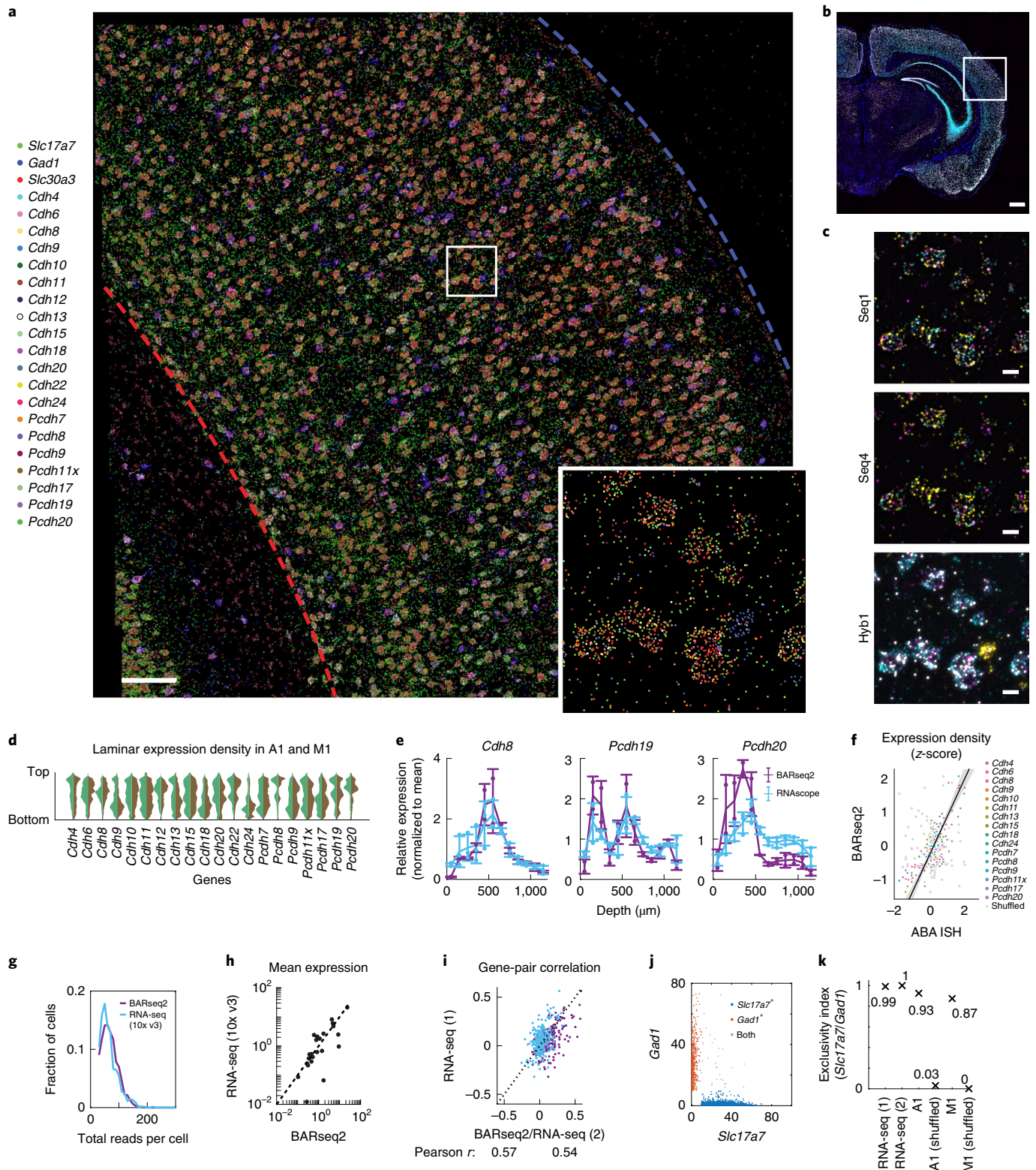
Further analyses showed that detection of mRNA by BARseq2 was specific. The mean expression of genes determined by BARseq2 was highly correlated with that determined by scRNA-seq using 10x v3 (Fig. 2h; Pearson correlation $r = 0.88$). A few outliers had substantially more counts in BARseq2 than in scRNA-seq, likely reflecting sampling differences across cell types, area-specific gene expression and differences in RNA accessibility in situ. For example, *Cdh6* expression observed by BARseq2 was 26 times that observed by scRNA-seq. This difference could be attributed to under-sampling of *Cdh6*-expressing pyramidal tract (PT) neurons in our scRNA-seq data⁶ and potentially variable sampling of neighboring cortical areas in which *Cdh6* is differentially expressed²². Furthermore, correlations between pairs of genes in single neurons determined by BARseq2 were consistent with scRNA-seq using 10x v3 to a similar extent as two independent 10x v3 experiments (Fig. 2i–k, Extended Data Fig. 5a,b and Methods). These results indicate that the single-cell gene expression patterns observed by BARseq2 were comparable to those of scRNA-seq.

We wondered if BARseq2 could detect more genes in parallel, and thus be potentially useful in associating projections with larger

Fig. 2 | Multiplexed detection of mRNAs using BARseq2. **a**, A representative image of rlonies in auditory cortex (from two slices sequenced). The top and the bottom of the cortex are indicated by the blue and red dashed lines, respectively. Scale bar, 100 μm. The inset shows a magnified view of the boxed area. **b**, Low-magnification image of the hybridization cycle, showing the location of the area imaged in **a**. Scale bar, 100 μm. **c**, Representative images of the indicated sequencing cycle and hybridization cycle of the boxed area in **a**. Scale bars, 10 μm. **d**, Violin plots showing the laminar distribution of cadherin expression in neuronal somata. Expression in auditory (green) and motor (brown) cortex is indicated. **e**, Laminar distribution of gene expression as detected by BARseq2 or FISH. Lines indicate means, error bars indicate s.d. values, and dots show individual data points. $n = 2$ slices for BARseq2 and $n = 3$ slices for FISH. **f**, Relative gene expression observed using BARseq2 and in Allen gene expression atlas. Each dot represents the expression of a gene in a 100-μm bin in laminar depth. Gray dots indicate the correlation between data randomized across laminar positions. A linear fit and 95% confidence intervals are shown by the diagonal line and the shaded area. $n = 2$ slices for BARseq2 and $n = 1$ slice for ABA ISH. **g**, Distribution of total read counts per cell in BARseq2 and scRNA-seq in auditory cortex. Only genes used in the panel detected by BARseq2 were included. **h**, Mean expression for each gene detected using BARseq2 or scRNA-seq. Each dot represents a gene. The dotted line indicates equal expression between BARseq2 and scRNA-seq. **i**, The correlations between pairs of genes observed in BARseq2 and scRNA-seq (purple dots), or in two scRNA-seq datasets (blue dots). **j**, Expression of *Slc17a7* and *Gad1* in single neurons. Neurons dominantly expressed *Slc17a7* (blue) or *Gad1* (red), or expressed both strongly (gray). **k**, Exclusivity indices (Methods) of *Slc17a7* and *Gad1* in neurons in two scRNA-seq datasets, BARseq2 in auditory or motor cortex, and shuffled BARseq2 data.

gene panels. Because BARseq2 imaging time scales logarithmically with the number of genes detected (Fig. 1d), the multiplexing capacity of BARseq2 is not limited by imaging time. Furthermore, targeting up to 65 genes did not significantly affect the detection sensitivity of each gene (Extended Data Fig. 5c and Methods). The detection of this 65-gene panel in motor cortex (Fig. 3a) allowed us to classify neurons to one of nine transcriptomic neuronal

types defined by scRNA-seq²³ (Fig. 3b, Methods and Extended Data Fig. 5d–h). Consistent with previous studies^{3,9}, these transcriptomic neuronal types displayed distinct laminar distributions (Fig. 3b,c and Methods) and cadherin expression (Fig. 3d). Most transcriptomic types were found in the expected layers with the notable exception of L5 PT and L6 IT Car3, which were seen in additional layers (for example, L2/3). These inaccuracies in cell typing likely



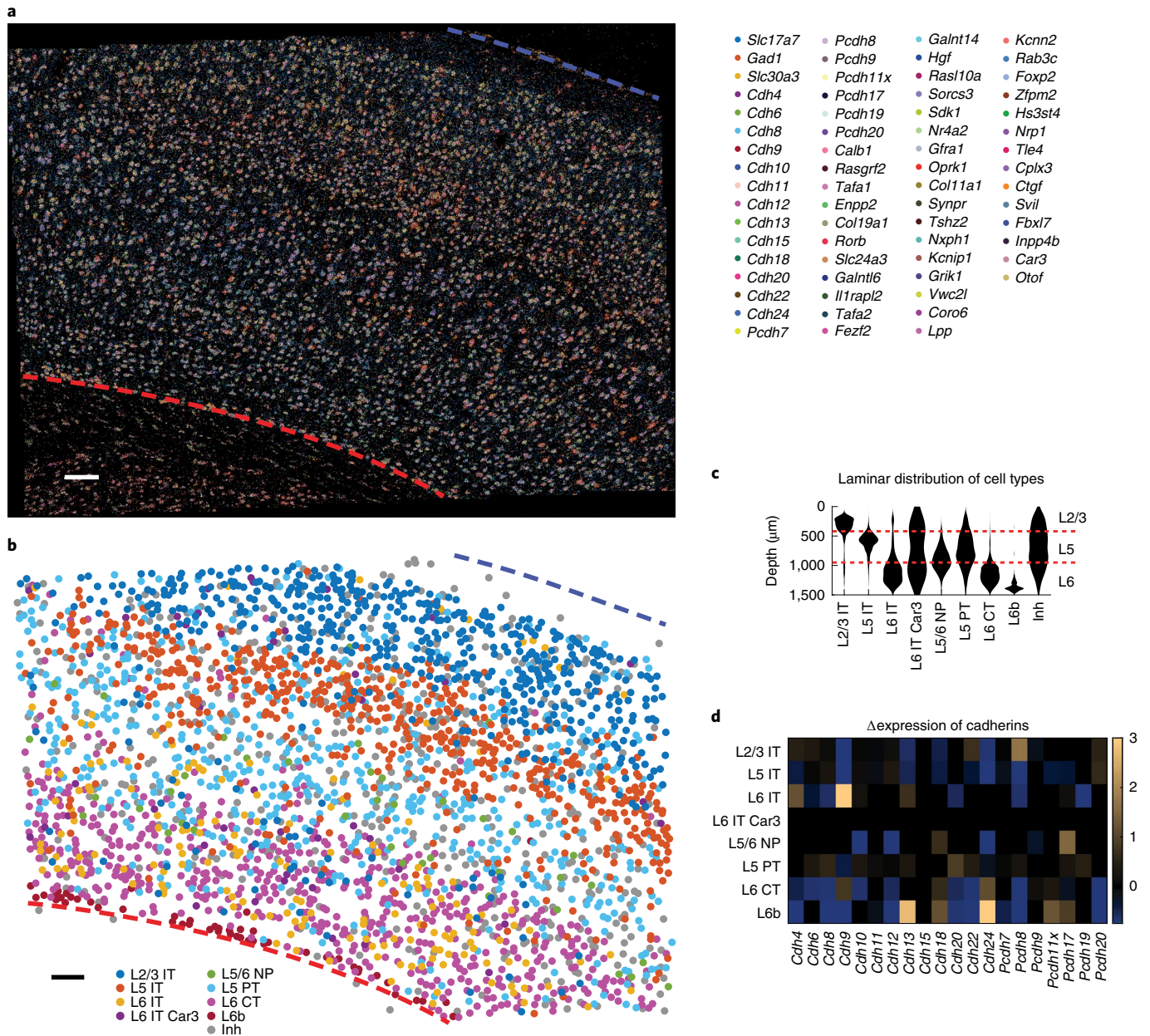


Fig. 3 | Cadherin expression across transcriptomic neuronal types in motor cortex. a, A representative image of rolonies in motor cortex (from four slices sequenced). mRNA identities are color-coded as indicated. The top and the bottom of the cortex are indicated by the blue and red dashed lines, respectively. Scale bar, 100 μm. **b**, Transcriptomic cell types called based on gene expression shown in **a**. **c**, Laminar distribution of transcriptomic neuronal types based on marker gene expression observed by BARseq2. Layer identities are shown on the right. NP, near-projecting. **d**, Differential expression of cadherins across transcriptomic neuronal types identified by BARseq2. Overexpression is indicated in yellow and underexpression is indicated in blue. Only differential expression that was statistically significant is shown. Statistical significance was determined using a two-tailed rank-sum test with Bonferroni correction for each gene between the indicated transcriptomic type and the expression of that gene across all other neuronal types.

resulted from suboptimal choice of marker genes (see Methods for a detailed discussion), and could potentially be improved in the future by optimizing the gene panels. These optimizations, however, were outside the scope of this study. These results demonstrate that BARseq2 can be applied to probe gene panels consisting of high dozens of genes, with minimal decrease in sensitivity and minimal increase in imaging time.

BARseq2 correlates gene expression to projections. Previous studies of the relationship between projection patterns and gene expression have largely focused on revealing the projection

patterns of transcriptomic neuronal types. Although this approach has identified some projection patterns biased in certain transcriptomic types^{6,8}, the diversity of projections in IT neurons remains largely unexplained by transcriptomic types^{3,6}. To further understand the relationship between gene expression and projections, we demonstrate an alternative approach that screens a targeted panel of genes for correlates of diverse projections. This approach relies on the ability of BARseq2 to interrogate both the expression of many genes and projections to many targets simultaneously, and thus would have been difficult to achieve using existing transcriptomic approaches that could only interrogate one or a small number of

projections (for example, Retro-seq^{3,9}) or barcoding-based projection mapping approaches that could only interrogate a small number of genes (for example, BARseq⁶).

As a proof-of-principle study, we examined long-range axonal projections and the expression of 20 cadherins, along with three marker genes, in motor cortex and auditory cortex in three mice. We optimized BARseq2 to detect both endogenous mRNAs and barcodes in the same barcoded neurons without compromising sensitivity (Extended Data Fig. 6a and Methods). In each barcoded cell, we segmented barcoded cell bodies using the barcode sequencing images (Fig. 4a). We then assigned colonies amplified from endogenous genes that overlapped with these pixels to the barcoded cells (Fig. 4a). This allowed us to map both projection patterns and gene expression (Fig. 4b) in the same neurons. We matched barcodes in these target sites to 3,164 well-segmented barcoded neurons (1,283 from auditory cortex and 1,881 from motor cortex) from 15 slices of auditory cortex and 16 slices of motor cortex, each with 10- μ m thickness. Of the barcoded neurons, 624 and 791 neurons had projections above the noise floor in auditory cortex and motor cortex, respectively. Most neurons (53% (329/624) in auditory and 89% (703/791) in motor cortex) projected to multiple brain areas. We then focused on 598 neurons in auditory cortex and 751 neurons in motor cortex, which also had sufficient endogenous mRNAs detected in each cell, for further analysis (Fig. 4c). These observations were largely consistent with previous BARseq experiments in auditory and motor cortex performed without assessing gene expression^{2,6}, confirming that the modifications for BARseq2 did not compromise projection mapping.

BARseq2 recapitulates known projection biases. Although BARseq2 can read out gene expression and projections in the same neurons, one might be concerned that barcoding neurons using Sindbis virus could disrupt gene expression²⁴. To determine the relationship between genes and projections, one would require that the gene–gene relationship in Sindbis-infected single neurons reflects that in noninfected neurons, and that any change in absolute gene expression level would have little effect. Reassuringly, previous reports have shown that the relationship among genes in single neurons is indeed largely preserved despite a reduction in the absolute expression of genes in Sindbis-infected cells^{6,25}. Furthermore, correlations between transcriptomic types and projections revealed in Sindbis-infected neurons were corroborated by other methods that did not require Sindbis infection^{6,26}. In agreement with these previous reports, we observed that the correlations between pairs of genes in the barcoded neurons were consistent with those in non-barcoded neurons despite an overall reduction in gene expression (Extended Data Fig. 6b–f and Methods). Therefore, the relationship between gene expression and projections resolved by BARseq2 likely reflects that in non-barcoded neurons.

To further test whether BARseq2 can capture the relationship between gene expression and projections, we asked if we could identify differences in projection patterns across transcriptomic neuronal types that could also be validated by previous studies and/or other experimental techniques. We performed these validation

analyses at three different levels of granularity. First, BARseq2 confirmed that most barcoded neurons with long-range projections were excitatory, not inhibitory; whereas about 8–9% of all barcoded neurons were inhibitory (100 of 1,047 in auditory cortex and 140 of 1,689 in motor cortex; Fig. 4d), only 7 of 240 (3%) inhibitory neurons (5 in auditory cortex and 2 in motor cortex) had detectable projections (Fig. 4e, Methods and Extended Data Fig. 6g,h). Second, BARseq2 identified many cadherins (8 for auditory cortex and 12 for motor cortex) that were differentially expressed across IT, PT and corticothalamic (CT) neurons²⁷ (Fig. 5a–d); the differential expression of these genes was consistent with the expression observed by scRNA-seq³ (Extended Data Fig. 7a and Methods). Finally, BARseq2 confirmed known biases in projection patterns across transcriptionally defined IT subtypes in auditory cortex (Extended Data Fig. 7b,c and Methods). Thus, BARseq2 recapitulated known projection differences across transcriptomic subtypes of IT neurons.

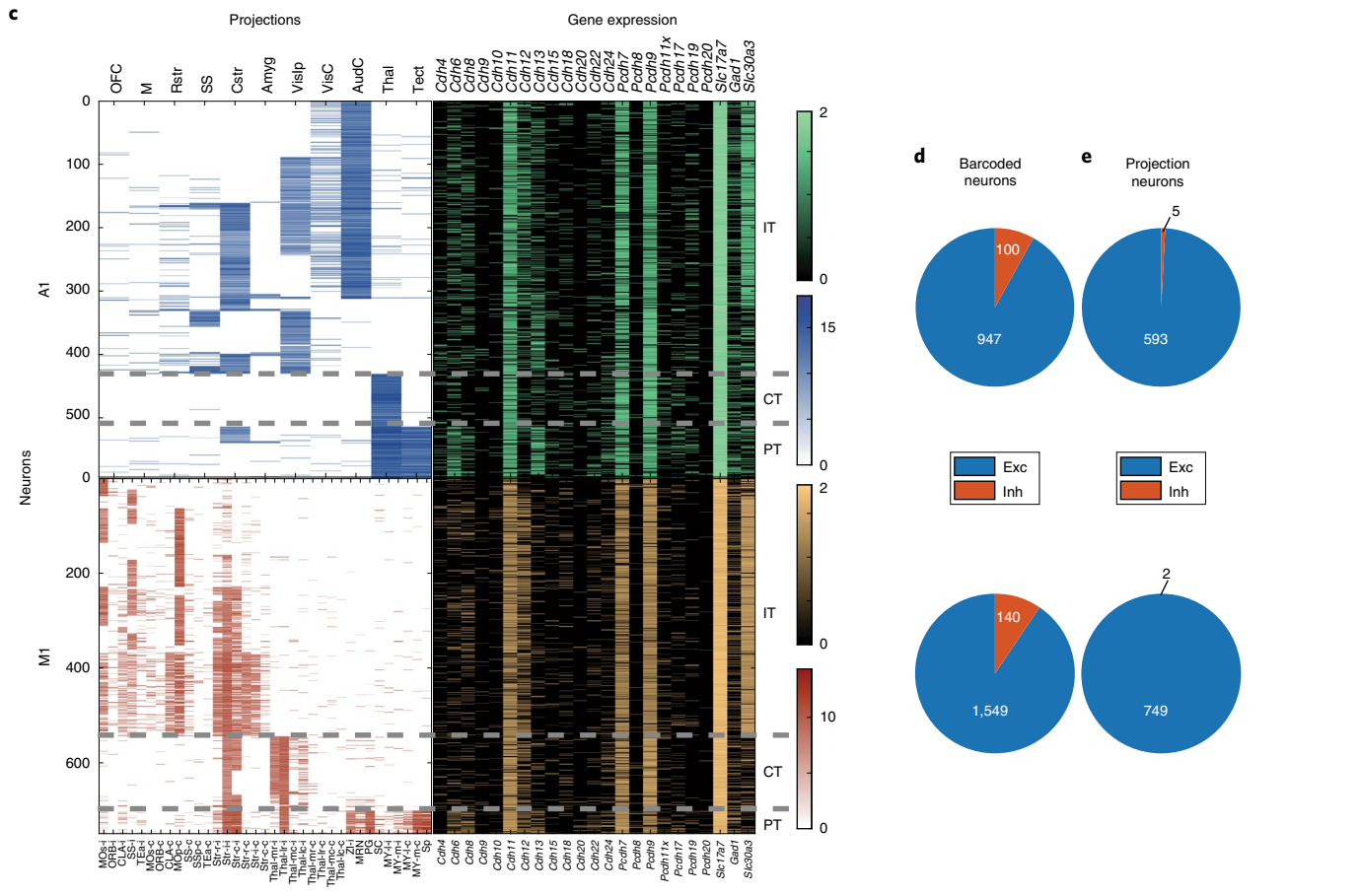
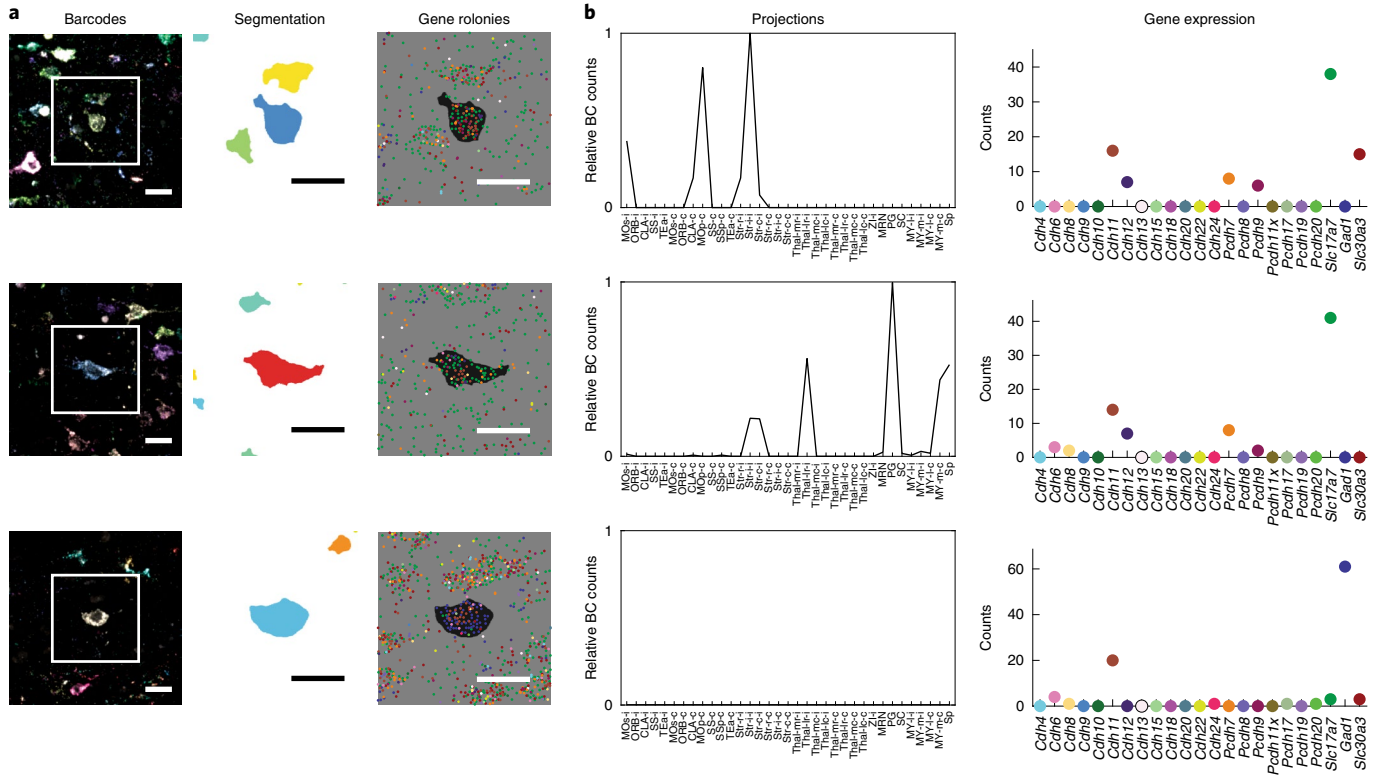
BARseq2 identifies cadherin correlates of IT projections. Having established that BARseq2 identified gene correlates of projections that were consistent with previous studies, we then asked whether cadherin expression correlates with projection patterns within the IT class of neurons. Although cadherins and other cell adhesion molecules are involved in projection specification and axonal growth during development^{16,28}, many take on other functions unrelated to projection specification in later developmental stages^{29,30}. In addition, other mechanisms such as axonal pruning could further shape the projection patterns of neurons independent of initial genetic programs. Therefore, any correlation between cadherins and projections is likely a remnant, or ‘echo,’ of the developmental program that initially specified projections, and may thus be weak and further obscured by gene expression associated with later developmental stages. To overcome the challenges of identifying potentially weak relationships between gene expression and projections, we used BARseq2 to identify correlations between projections and cadherins using a module-based strategy inspired by similar approaches in transcriptomics³¹. Projection modules and gene modules average over the noise in the measurement of individual projections and genes, respectively, and are thus easier to detect when there is considerable biological and/or technical noise in the measurements. This approach requires knowing the projections to many brain areas from individual neurons, a unique advantage of barcoding-based projection mapping techniques (that is, BARseq and BARseq2) compared to retrograde labeling-based approaches^{3,9}. Next, we identify modest associations between cadherin expression and projections in IT neurons, including several associated pairs of cadherins/projections that were shared across cortical areas.

The projections of an IT neuron to its targets are not random. Rather, in both auditory cortex and motor cortex, these projections are organized and show statistical regularities that can be uncovered within the large datasets obtained by BARseq^{2,6} (Fig. 6a). For example, neurons in the auditory cortex that projected to the somatosensory cortex were also more likely to project to the ipsilateral visual cortex, but not the contralateral auditory cortex. To

Fig. 4 | Correlating gene expression to projections using BARseq2. **a**, False-colored barcode sequencing images, soma segmentations and gene colonies of three representative neurons from the motor cortex. The segmentation and gene colony images correspond to the white squared area in the barcode images. In the gene colony images, the areas corresponding to the soma segmentations of the target neurons are in black. Scale bars, 20 μ m. **b**, Projections and gene expression of the target neurons shown in **a**. The dots indicating gene expression use the same color coding as in **a**. The neurons shown in the first two rows are excitatory projection neurons, whereas the neuron shown in the bottom row is an inhibitory neuron without projections. See Supplementary Table 2 for the brain areas corresponding to each abbreviated target area. BC, barcode. **c**, Projections and gene expression of neurons in auditory cortex (A1) and motor cortex (M1). Each row represents a barcoded projection neuron. Both projections and gene expression are shown in log scale. Major projection neuron classes determined by projection patterns are indicated on the right. **d,e**, The number of excitatory (exc) or inhibitory (inh) neurons in all barcoded neurons (**d**) or barcoded projection neurons (**e**). Neurons in auditory cortex are shown in the top row and those in motor cortex are shown in the bottom row.

exploit these correlations, we used nonnegative matrix factorization (NMF)³² to represent the projection pattern of each neuron as the sum of several ‘projection modules.’ (NMF is an algorithm related

to principal-component analysis, but imposes the added constraint that projections are nonnegative). Each of these modules (six modules for the motor cortex and three for the auditory cortex; Fig. 6b)



consisted of subsets of projections that were likely to co-occur. We named these modules by the main projections (cortex (CTX) or striatum (STR)) followed by the side of the projection (ipsilateral (I) or contralateral (C)). For some modules, we further indicated that the projections were to the caudal part of the structure by prefixing with 'C' (for example, CSTR-I or CCTX-I). A small number of projection modules could explain most of the variance in projections (three modules and six modules explained 84% and 87% of the variance in projections to nine areas in auditory cortex and 18 areas in motor cortex that IT neurons project to, respectively; Fig. 6c).

Because both the projection patterns of neurons^{2,27} and their transcriptomic types^{3,9} are well correlated with laminae, we first asked how well cadherins explained the diversity of projections in IT neurons compared to the laminar positions of neurons (Methods). Although most cadherins had no predictive power on the projection modules, some individual cadherins could explain a substantial fraction of the variance in projections compared to that explained by the laminar positions of neurons (Extended Data Fig. 8). For example, *Cdh13* and *Pcdh7* explained $6.0\% \pm 0.3\%$ and $7.0\% \pm 0.3\%$ (mean \pm s.d.) of the variations in CTX-C in auditory cortex, compared to $19.4 \pm 0.3\%$ (mean \pm s.d.) explained by the laminar positions of neurons. Strikingly, *Pcdh19* and *Pcdh7* were predictive of CSTR-I in auditory cortex, whereas the laminar positions were not. These results indicate that some but not all cadherins were modestly predictive of projections, and that the predictive power of these cadherins could be comparable in magnitude to the laminar positions of neurons, one of the strongest known predictors of projection patterns.

To further understand how cadherin expression relates to projections, we examined how it covaried with projection modules (Supplementary Fig. 1). Interestingly, the expression of several cadherins covaried with similar projection modules in both cortical areas. For example, auditory cortex neurons expressing *Pcdh19* were stronger in the CSTR-I projection module than those not expressing *Pcdh19* (Fig. 6d; $P = 5 \times 10^{-4}$ comparing the CSTR-I module in neurons with ($n = 83$) or without ($n = 346$) *Pcdh19* expression using rank-sum test); the same association between *Pcdh19* and the CSTR-I projection module was also seen in motor cortex (Fig. 6d; $P = 4 \times 10^{-6}$ using rank-sum test, $n = 31$ for *Pcdh19*⁺ neurons and $n = 512$ for *Pcdh19*⁻ neurons). Similarly, *Cdh8* was correlated with the CTX-I module and *Cdh12* was correlated with the CTX-C module (Fig. 6e; false discovery rate (FDR) < 0.1) in both auditory and motor cortex. These correlations were independently validated by retrograde tracing using cholera toxin subunit B (CTB) and FISH (Extended Data Fig. 9a–e and Methods). *Pcdh19*, together with *Cdh8* or *Cdh11*, correlated with CTX-I or CSTR-I modules, respectively, in motor cortex (Fig. 6e and Extended Data Fig. 8), consistent with a potential combinatorial nature of cadherin correlates of projections. Although the correlations between individual cadherins and projections were relatively modest, our observations that the same cadherins correlated with similar projection modules in both areas suggest that a common molecular logic might underscore the

organization of projections across cortical areas beyond class-level divisions.

Analyses based on the expression of single genes suffer from biological and technical noise of gene expression in single neurons. We reasoned that the correlations among genes might allow us to identify additional relationships between gene expression and projections that were missed by analyzing each gene separately. This ability to leverage the relationship among genes represents an advantage of BARseq2 over the original BARseq because of the improved capacity of BARseq2 for multiplexed gene detection. To exploit the correlations among genes, we grouped 16 cadherins into three meta-analytic coexpression modules based on seven scRNA-seq datasets of IT neurons in motor cortex (Fig. 7a and Extended Data Fig. 10a,b)²³. To obtain the modules, we followed the rank-based network aggregation procedure defined by Ballouz et al.³³ and Crow et al.³⁴ to combine the seven dataset-specific gene–gene coexpression networks into an aggregated network, and then grouped together genes showing consistent excess correlation using the dynamic tree-cutting algorithm³¹. Two coexpressed modules were associated with projections: module 1 was associated with contralateral striatal projections (STR-C projection module), and module 2 was associated with ipsilateral caudal striatal projections (CSTR-I; Fig. 7b,c and Extended Data Fig. 10c,d). These associations between the coexpression modules and projections were consistent with, but stronger than, associations between individual genes contained in each module and the same projections (Extended Data Fig. 10e). Interestingly, these coexpression modules were enriched in multiple transcriptomic subtypes of IT neurons, but these transcriptomic subtypes were found in multiple branches of the transcriptomic taxonomy (Fig. 7d and Extended Data Fig. 10f). For example, module 1 was associated with transcriptomic subtypes of IT neurons in L2/3, L5 and L6. This result is consistent with previous observations^{3,6} that first-tier transcriptomic subtypes of IT neurons (that is, subtypes of the highest level in the transcriptomic taxonomy within the IT class) appeared to share projection patterns, and further raises the possibility that transcriptomic taxonomy does not necessarily capture differences in projections. Taken together, our finding that projections correlate with cadherin coexpression modules independent of transcriptomic subtypes demonstrates that BARseq2 can reveal intricate relationships between gene expression and projection patterns.

Discussion

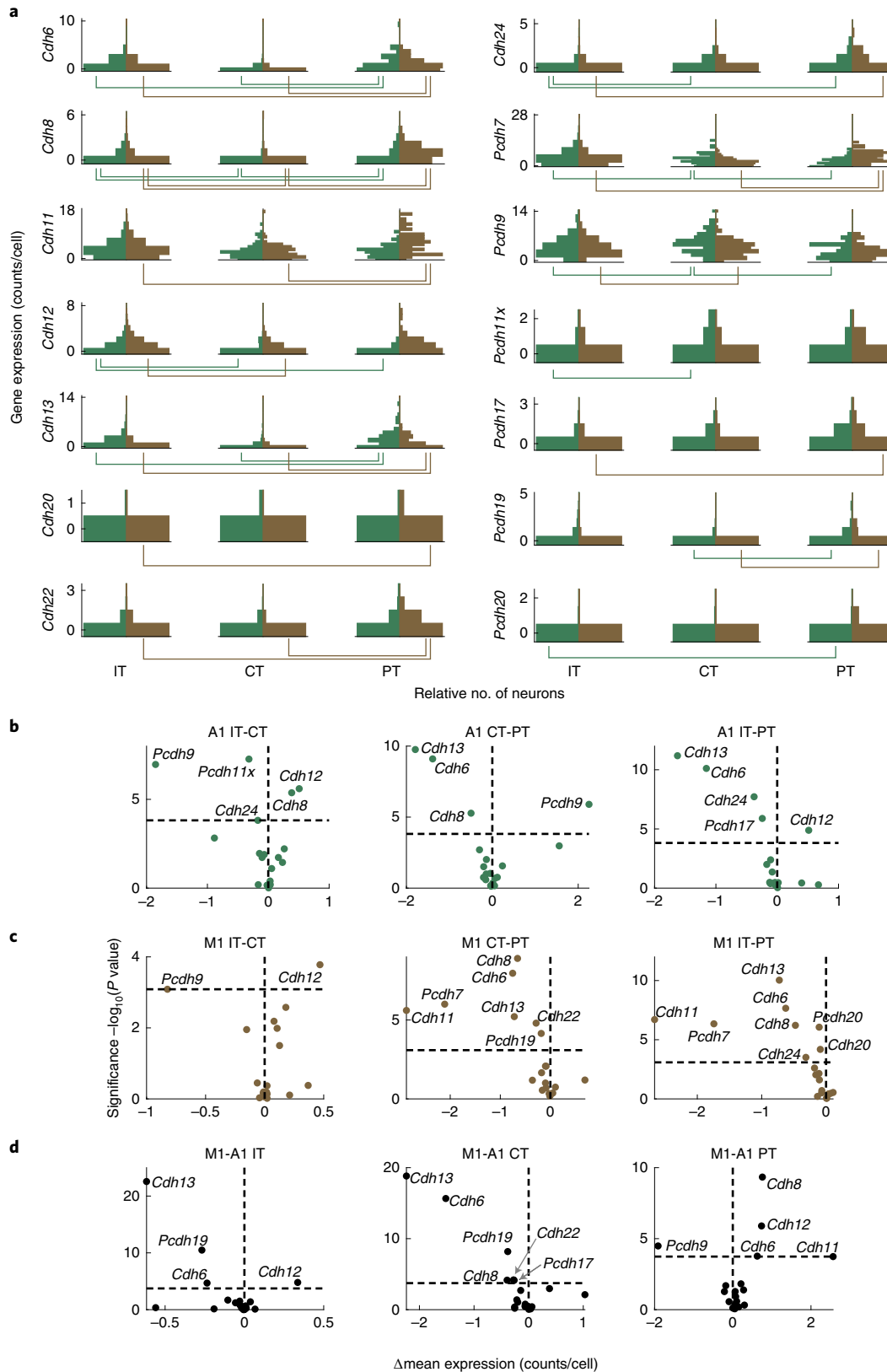
BARseq2 combines high-throughput mapping of projections to many brain areas with multiplexed detection of gene expression at single-cell resolution. Because BARseq2 is high throughput, we are able to correlate gene expression and projection patterns of thousands of individual neurons in a single experiment, and thereby achieve statistical power that would be challenging to obtain using other single-cell techniques. By applying BARseq2 to two distant cortical areas—primary motor and auditory cortex—in the adult mouse, we identified cadherin correlates of diverse projections. Our results suggest that BARseq2 provides a path to discovering

Fig. 5 | Differential cadherin expression across major classes and cortical areas. **a**, Vertical histograms of the expression (raw counts per cell) of cadherins that were differentially expressed across major classes in either auditory or motor cortex. *y* axes indicate gene expression level (counts per cell) and *x* axes indicate number of neurons at that expression level. The numbers of neurons are normalized across plots so that the bins with the maximum number of neurons have equal bar lengths. In each plot, gene expression in auditory cortex (green) is shown on the left, and gene expression in motor cortex (brown) is shown on the right. Lines beneath each plot indicate pairs of major classes with different expression of the gene (FDR < 0.05). **b,c**, Volcano plots of cadherins that were differentially expressed across pairs of major classes in auditory cortex (**b**) or motor cortex (**c**). *y* axes indicate significance and *x* axes indicate effect size. The horizontal dashed lines indicate significance level for FDR < 0.05 , and the vertical dashed lines indicate equal expression. **d**, Volcano plots of cadherins that were differentially expressed across auditory and motor cortex in the indicated major classes. *y* axes indicate significance and *x* axes indicate effect size. Gene identities for points close together are noted with gray arrows for clarity. The horizontal dashed lines indicate significance level for FDR < 0.05 , and the vertical dashed lines indicate equal expression. For all panels, *P* values were calculated using two-tailed rank-sum tests.

the general organization of gene expression and projections that are shared across the cortex.

High-throughput and multiplexed gene detection by BARseq2. To correlate panels of genes to projections, we designed BARseq2

to detect gene expression with high throughput, for multiplexing to dozens of genes, to have sufficient sensitivity, and be compatible with barcoding-based projection mapping. To satisfy these needs, we based BARseq2 on padlock probe-based approaches^{10,11}. With additional optimizations for sensitivity, sequencing readout



and compatibility with barcode sequencing, we successfully used BARseq2 to identify gene correlates of projections.

One of the critical requirements for BARseq2 is high throughput when reading out many genes. Through strong amplification of mRNAs, combinatorial coding and robust readout using Illumina sequencing chemistry^{6,35}, BARseq2 achieves fast imaging at low optical resolution compared to many other imaging-based spatial transcriptomic methods^{14,36}. Further optimizations, including computational approaches for resolving spatially mixed colonies³⁷, have the potential to increase imaging throughput even further. Although the gene multiplexing capacity of BARseq2 may ultimately be limited by other physical constraints, such as crowding of colonies and reduced detection sensitivity, these factors are unlikely to be limiting when multiplexing to dozens to hundreds of genes¹¹.

Another critical optimization was increasing the low sensitivity that early versions of the padlock probe-based technique was susceptible to, unless special and expensive primers were used¹⁰. Inspired by other spatial transcriptomic methods, we and others¹¹ have found that tiling target genes with multiple probes could greatly improve the sensitivity. This design allowed variable sensitivity for different experimental purposes. Although in the present work we identified cadherin correlates of projections using only a modest number of probes per gene to achieve sensitivity similar to scRNA-seq using 10x Genomics v3, the sensitivity of BARseq2 can be considerably higher when more probes are used (Fig. 1e). This high and tunable sensitivity, combined with the fact that the gene multiplexing capacity of BARseq2 is not limited by imaging time, opens potential application of BARseq2 to a wide range of questions that require high-throughput interrogation of gene expression in situ.

BARseq2 reveals gene correlates of projections. BARseq2 exploits the high-throughput axonal projection mapping that BARseq offers to identify gene correlates of diverse projections. BARseq has sensitivity comparable to single-neuron tracing⁵. Although the spatial resolution of BARseq for projections is lower than that of conventional single-neuron tracing, it offers throughput that is several orders of magnitude higher than the state-of-the-art single-cell tracing techniques¹². This high throughput allows BARseq to reveal higher-order statistical structure in projection patterns that would have been difficult to observe using existing techniques, such as single-cell tracing^{5,6}. The increased statistical power of BARseq, obtained at the cost of some spatial resolution, is reminiscent of different clustering power across scRNA-seq techniques of varying throughput and read depth^{23,38}. The high throughput of BARseq thereby provides a powerful asset for investigating the organization of projection patterns and their relationship to gene expression.

BARseq2 enables simultaneous measurement of multiplexed gene expression and axonal projections to many brain areas, at single-neuron resolution and at a scale that would be difficult to achieve with other approaches. For example, *Cre*-dependent

labeling allows interrogation of the gene expression and projection patterns of a genetically defined subpopulation of neurons⁵. However, this approach lacks cellular resolution, is limited by the availability of *Cre* lines, and requires that a neuronal population of interest be specifically distinguished by the expression of one or two genes. The combination of single-cell transcriptomic techniques with retrograde labeling does provide cellular resolution, but can only interrogate projections to one or at most a small number of brain areas at a time^{3,9}. The inability to interrogate projections to many brain areas from the same neuron would miss higher-order statistical structures in projections, which are nonrandom⁵ and provide additional information regarding other properties of the neurons, such as laminar position and gene expression^{2,6}. The projections of individual neurons to multiple brain areas can be obtained using multiplexed single-cell tracing¹, but the throughput of these methods remains relatively low. Moreover, many advanced single-cell tracing techniques require special sample processing that hinders multiplexed interrogation of gene expression in the same sample. The throughput of single-cell projection mapping was addressed by the original BARseq⁶, but the small number of genes (up to three) that could be co-interrogated with projections limited its use in identifying the general relationship between gene expression and projections. BARseq2 thus addresses limitations of existing techniques and provides a powerful approach for probing the relationships between gene expression and projection patterns.

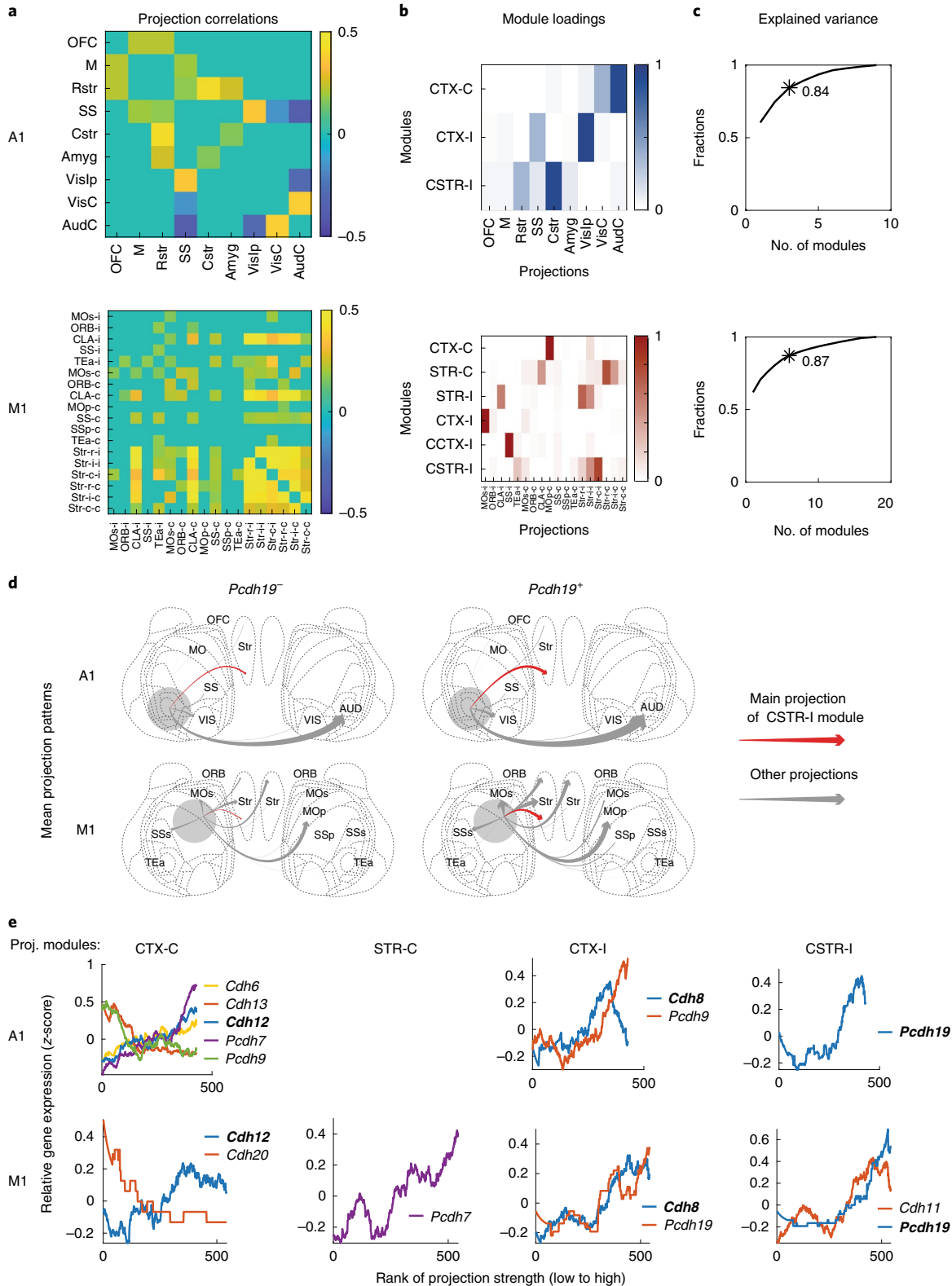
Cadherins correlate with diverse projections of IT neurons. As a proof-of-principle study, we used BARseq2 to identify several cadherins that correlate with homologous IT projections in both auditory and motor cortex, two spatially and transcriptomically distant areas with distinct cortical and subcortical projection targets. In addition, cadherin coexpression modules that correlated with projections were associated with multiple branches of the transcriptomic taxonomy. This type of correlation between neuronal connectivity and variations in gene expression independent of transcriptomic types is not unique to the cortex and has previously been observed in other brain areas, such as the hippocampus³⁹. Therefore, our findings are consistent with the hypothesis that a shared cell adhesion molecule code might underlie the diversity of cortical projections independent of transcriptomic types^{18,39}.

Even though the power of some cadherins to predict projections was comparable in magnitude to that of laminar position, a strong predictor of projection patterns, these cadherins could only explain a small fraction of the overall variance in projections. This noisy association between cadherin expression and projection patterns contrasts with the known roles of cadherins in specifying neuronal connectivity in the cortex and other circuits^{20,40}, but the relatively small magnitude of these associations is not surprising for a few reasons. First, gene expression programs and signaling cues needed for specifying projections are usually transient in development⁴¹, so it is likely that these cadherins only represent the remnants of

Fig. 6 | Cadherins correlate with diverse projections of IT neurons. **a**, Pearson correlation of projections to different brain areas in IT neurons of auditory cortex (A1) or motor cortex (M1). Only significant correlations are shown. OFC, orbitofrontal cortex; M, motor cortex; Rstr, rostral striatum; SS, somatosensory cortex; Cstr, caudal striatum; Amyg, amygdala; VisIp, ipsilateral visual cortex; VisC, contralateral visual cortex; AudC, contralateral auditory cortex. **b**, Projection modules of IT neurons in auditory cortex (top) or motor cortex (bottom). Each row represents a projection module. Columns indicate projections to different brain areas. **c**, The fractions of variance explained by different numbers of projection modules in auditory cortex (top) and motor cortex (bottom). The numbers of projection modules that correspond to those in **b** are labeled with an asterisk with the fraction of variance explained indicated. **d**, Mean projection patterns of neurons in A1 and M1 with or without *Pcdh19* expression. The thickness of arrows indicates projection strength (barcode counts). Red arrows indicate projections that correspond to the strongest projection in the CSTR-I projection modules. ORB, orbitofrontal cortex; MOs, secondary motor cortex; MOp, primary motor cortex; SSp, primary somatosensory cortex; SSs, secondary somatosensory cortex; TEa, temporal association cortex. **e**, The expression of cadherins (y axes) that were rank correlated with the indicated projection modules in auditory cortex and motor cortex. Neurons (x axes) were sorted by the strengths of the indicated projection modules. Only genes that were significantly correlated with projection modules are shown (FDR < 0.1 using two-tailed rank-sum tests). Genes that were correlated with the same projection modules in both areas are shown in bold.

a common developmental program that establish projections⁴², or may be needed for ongoing functions or maintenance of projections. Second, non-cadherin cell adhesion molecules (for example, IgCAMs^{43,44}) and other cell-surface molecules (for example, plexins, semaphorins⁴⁵ and teneurins⁴⁶) are also involved in specifying projections, so cadherins likely only represent a fraction of the molecular programs that specify projections. Finally, cortical projections

undergo extensive activity-dependent modifications after the initial specification, so the overall diversity in cortical projections is likely much higher than that produced by the initial molecular program. These possibilities can be better resolved by applying BARseq2 to reveal gene expression in both the projection neurons and the areas they project to during development, in combination with perturbation experiments. BARseq2 thus provides a path to discovering



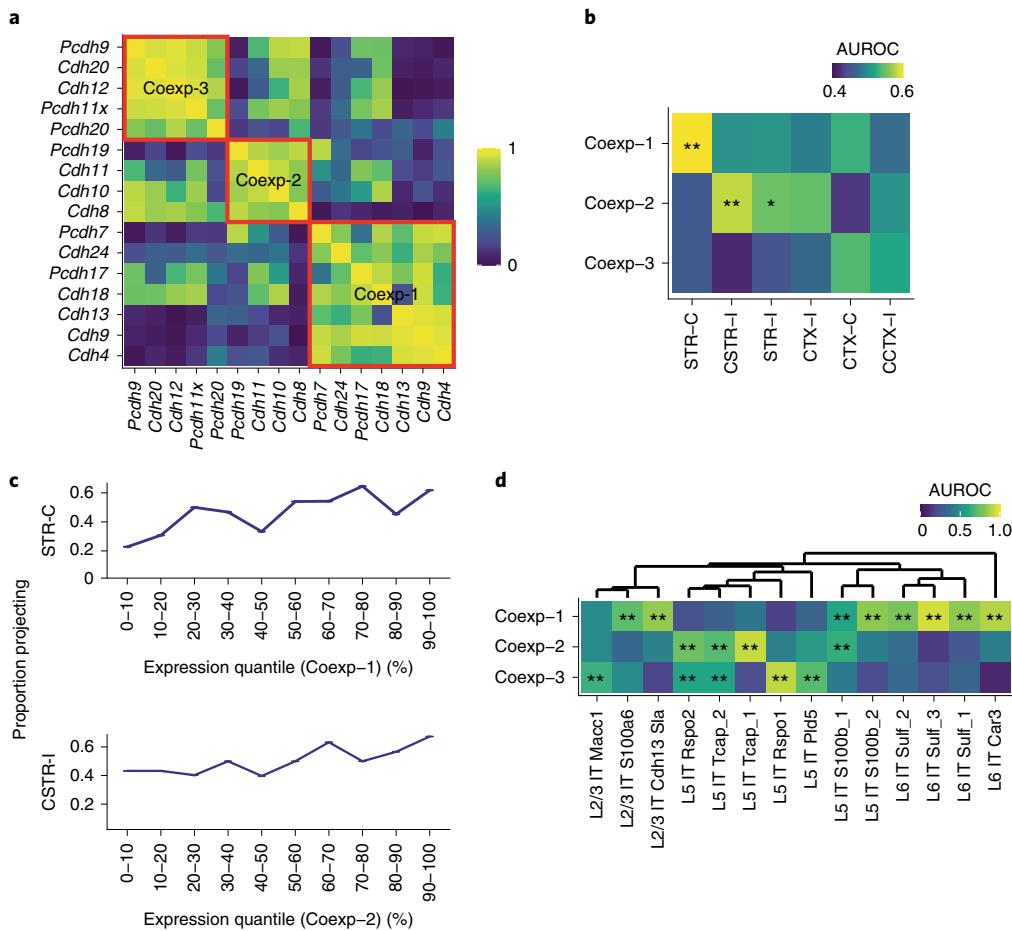


Fig. 7 | Gene coexpression modules correlate with diverse projections of IT neurons. **a**, Correlation among cadherins as identified using scRNA-seq in IT neurons in motor cortex²³. Three coexpression modules are marked by red squares. Cadherins that did not belong to any module are not shown. **b**, Association between cadherin coexpression modules and projection modules (AUROC, area under the receiver-operator characteristic curve). Significant associations are marked by asterisks (*FDR < 0.1, **FDR < 0.05). **c**, Fractions of neurons with the indicated projection modules as a function of coexpression module expression. Neurons are binned by gene module quantiles as indicated. **d**, Association of the three coexpression modules in transcriptomic IT neurons in the single-cell SmartSeq dataset (AUROC, significance shown as in **b**).

the myriad of genetic programs that specify and/or correlate with long-range projections in both developing and mature animals.

BARseq2 builds a unified description of neuronal diversity.

Neuronal barcoding was originally proposed as a method for untangling circuit connectivity at synaptic resolution^{47,48}. Solving neuronal connectivity with barcode sequencing not only has the potential to achieve high-throughput and single-cell resolution by exploiting advances in sequencing technology, but also provides a path to integrate measurements of multiple neuronal properties in single neurons—toward the ‘Rosetta brain’⁴⁹. BARseq2 is a step toward this goal. Although BARseq2 currently only resolves projections at relatively low spatial resolution (brain areas, that is hundreds of microns), this limitation can be addressed in the future by using in situ sequencing to read out axonal barcodes (Yuan et al., unpublished data), which would resolve axonal projections at sub-cellular spatial resolution. Further combining in situ sequencing of axonal barcodes with synaptic labeling, expansion microscopy and/or transsynaptic viral labeling could yield information regarding the synaptic connectivity of neurons. Because BARseq2 integrates neuronal properties using spatial information, it is potentially compatible with other in situ assays, such as immunohistochemistry, two-photon calcium imaging and dendritic morphological reconstruction. By spatially correlating various neuronal properties in

single neurons, BARseq2 represents a feasible path toward achieving a comprehensive description of neuronal circuits.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-021-00842-4>.

Received: 25 August 2020; Accepted: 19 March 2021;

Published online: 10 May 2021

References

1. Winnubst, J. et al. Reconstruction of 1,000 projection neurons reveals new cell types and organization of long-range connectivity in the mouse brain. *Cell* **179**, 268–281 (2019).
2. Muñoz-Castañeda, R. et al. Cellular anatomy of the mouse primary motor cortex. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.10.02.323154> (2020).
3. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
4. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 (2018).
5. Han, Y. et al. The logic of single-cell projections from visual cortex. *Nature* **556**, 51–56 (2018).

6. Chen, X. et al. High-throughput mapping of long-range neuronal projection using in situ sequencing. *Cell* **179**, 772–786 (2019).
7. Kim, D. W. et al. Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell* **179**, 713–728 (2019).
8. Economo, M. N. et al. Distinct descending motor cortex pathways and their roles in movement. *Nature* **563**, 79–84 (2018).
9. Zhang, M. et al. Molecular, spatial and projection diversity of neurons in primary motor cortex revealed by in situ single-cell transcriptomics. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.06.04.105700> (2020).
10. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
11. Qian, X. et al. Probabilistic cell typing enables fine mapping of closely related cell types in situ. *Nat. Methods* **17**, 101–106 (2020).
12. Kebschull, J. M. et al. High-throughput mapping of single-neuron projections by sequencing of barcoded RNA. *Neuron* **91**, 975–987 (2016).
13. Huang, L. et al. BRICseq bridges brain-wide interregional connectivity to neural activity and gene expression in single animals. *Cell* **182**, 177–188 (2020).
14. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
15. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
16. Hayano, Y. et al. The role of T-cadherin in axonal pathway formation in neocortical circuits. *Development* **141**, 4784–4793 (2014).
17. Friedman, L. G. et al. Cadherin-8 expression, synaptic localization, and molecular control of neuronal form in prefrontal corticostriatal circuits. *J. Comp. Neurol.* **523**, 75–92 (2015).
18. Paul, A. et al. Transcriptional architecture of synaptic communication delineates GABAergic neuron identity. *Cell* **171**, 522–539 (2017).
19. Matsunaga, E., Nambu, S., Oka, M. & Iriki, A. Complex and dynamic expression of cadherins in the embryonic marmoset cerebral cortex. *Dev. Growth Differ.* **57**, 474–483 (2015).
20. Redies, C. Cadherins and the formation of neural circuitry in the vertebrate CNS. *Cell Tissue Res.* **290**, 405–413 (1997).
21. Lein, E. S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
22. Terakawa, Y. W., Inoue, Y. U., Asami, J., Hoshino, M. & Inoue, T. A sharp cadherin-6 gene expression boundary in the developing mouse cortical plate demarcates the future functional areal border. *Cereb. Cortex* **23**, 2293–2308 (2013).
23. Yao, Z. et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.02.29.970558> (2020).
24. Fros, J. J. & Pijlman, G. P. Alphavirus infection: Host cell shut-off and inhibition of antiviral responses. *Viruses* <https://doi.org/10.3390/v8060166> (2016).
25. Klingler, E. et al. Single-cell molecular connectomics of intracortically projecting neurons. Preprint at *bioRxiv* <https://doi.org/10.1101/378760> (2018).
26. Wang, Y. et al. Complete single-neuron reconstruction reveals morphological diversity in molecularly defined claustral and cortical neuron types. Preprint at *bioRxiv* <https://doi.org/10.1101/675280> (2019).
27. Harris, K. D. & Shepherd, G. M. The neocortical circuit: themes and variations. *Nat. Neurosci.* **18**, 170–181 (2015).
28. Duan, X., Krishnaswamy, A., De la Huerta, I. & Sanes, J. R. Type II cadherins guide assembly of a direction-selective retinal circuit. *Cell* **158**, 793–807 (2014).
29. Friedman, L. G., Benson, D. L. & Huntley, G. W. Cadherin-based transsynaptic networks in establishing and modifying neural connectivity. *Curr. Top. Dev. Biol.* **112**, 415–465 (2015).
30. Jontes, J. D. The cadherin superfamily in neural circuit assembly. *Cold Spring Harb. Perspect. Biol.* **10**, a029306 (2018).
31. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
32. Lee, D. D. & Seung, H. S. Learning the parts of objects by nonnegative matrix factorization. *Nature* **401**, 788–791 (1999).
33. Ballouz, S., Verleyen, W. & Gillis, J. Guidance for RNA-seq coexpression network construction and analysis: safety in numbers. *Bioinformatics* **31**, 2123–2130 (2015).
34. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Exploiting single-cell expression to characterize coexpression replicability. *Genome Biol.* **17**, 101 (2016).
35. Chen, X., Sun, Y. C., Church, G. M., Lee, J. H. & Zador, A. M. Efficient in situ barcode sequencing using padlock probe-based BaristaSeq. *Nucleic Acids Res.* **46**, e22 (2018).
36. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342–357 (2016).
37. Chen, S. et al. BARcode DEMixing through Non-negative Spatial Regression (BarDensr). *PLoS Comput. Biol.* **17**, e1008256 (2021).
38. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
39. Harris, K. D. et al. Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS Biol.* **16**, e2006387 (2018).
40. Duan, X. et al. Cadherin combinations recruit dendrites of distinct retinal neurons to a shared interneuronal scaffold. *Neuron* **99**, 1145–1154 (2018).
41. Li, H. et al. Classifying *Drosophila* olfactory projection neuron subtypes by single-cell RNA sequencing. *Cell* **171**, 1206–1220 (2017).
42. Custo Greig, L. F., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat. Rev. Neurosci.* **14**, 755–769 (2013).
43. Bagri, A. et al. Slit proteins prevent midline crossing and determine the dorsoventral position of major axonal pathways in the mammalian forebrain. *Neuron* **33**, 233–248 (2002).
44. Shu, T., Sundaresan, V., McCarthy, M. M. & Richards, L. J. Slit2 guides both precrossing and postcrossing callosal axons at the midline in vivo. *J. Neurosci.* **23**, 8176–8184 (2003).
45. Yoshida, Y. Semaphorin signaling in vertebrate neural circuit assembly. *Front. Mol. Neurosci.* **5**, 71 (2012).
46. Berns, D. S., DeNardo, L. A., Pederick, D. T. & Luo, L. Teneurin-3 controls topographic circuit assembly in the hippocampus. *Nature* **554**, 328–333 (2018).
47. Zador, A. M. et al. Sequencing the connectome. *PLoS Biol.* **10**, e1001411 (2012).
48. Peikon, I. D. et al. Using high-throughput barcode sequencing to efficiently map connectomes. *Nucleic Acids Res.* **45**, e115 (2017).
49. Marblestone, A. H., et al. Rosetta brains: a strategy for molecularly-annotated connectomics. Preprint at *arXiv* <https://arxiv.org/abs/1404.5103> (2014).
50. Eng, C. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

Animal processing and tissue preparation. All animal procedures were carried out in accordance with the Institutional Animal Care and Use Committee (protocol no. 19-16-10-07-03-00-4) at Cold Spring Harbor Laboratory. The animals were housed at maximum of five in a cage on a 12-h on/12-h off light cycle. The temperature in the facility was kept at 22°C with a range not exceeding 20.5°C to 26°C. Humidity was maintained at around 45–55%, not exceeding a range of 30–70%. A list of animals used is provided in Supplementary Table 1.

For samples used for only endogenous mRNA detection, 8- to 10-week-old male C57BL/6 mice were anesthetized and decapitated. We immediately embedded the brain in optimal cutting temperature (OCT) compound in a 22-mm² cryomold and snap froze the tissue in an isopentane bath submerged in liquid nitrogen. Sections were cut into 10- μ m-thick slices on Superfrost Plus Gold Slides (Electron Microscopy Sciences). Unlike in the original BARseq, the sections were directly melted onto slides without the use of a tape-transfer system. This change in mounting methods allowed increased efficiency in gene detection. The slides were stored at –80°C until use.

For BARseq2 samples, 8- to 10-week-old male C57BL/6 mice were injected as indicated in Supplementary Table 1. After 24 h, we anesthetized and decapitated the animal, punched out the injection site and snap froze the rest of the brain on a razor blade on dry ice for conventional MAPseq⁶. The injection site was embedded, cryosectioned and stored as described above.

To prepare samples for BARseq2 experiments, we immersed slides from –80°C instantly into freshly made 4% paraformaldehyde (10-ml vials of 20% PFA; Electron Microscopy Sciences) in PBS for 30 min at room temperature. We washed the samples in PBS for 5 min before installing HybriWell-FL chambers (22 mm \times 22 mm \times 0.25 mm; Grace Bio-Labs) for subsequent reactions on the samples. We then dehydrated the samples in 70%, 85% and 100% ethanol for 5 min each, followed by washing in 100% ethanol for at least 1 h at 4°C. Finally, we rehydrated the samples in PBST (0.5% Tween-20 in PBS).

For retrograde labeling experiments, we prepared 1.0 mg ml⁻¹ of CTB in PBS from 100 μ g for injections (see Supplementary Table 1 for a list of animals and coordinates used). We perfused the animals with fresh 4% PFA 96 h after injection, post-fixed for 24 h in 4% PFA, and cryoprotected in 10% sucrose in PBS for 12 h, 20% sucrose in PBS for 12 h and 30% sucrose in PBS for 12 h. The brain was then frozen in OCT and cryosectioned into 20- μ m slices using a tape-transfer system.

BARseq2 detection of endogenous genes. We prepared a master mix of reverse transcription primers at 0.5 μ M each for all target mRNAs. For volumes exceeding the amount required for reverse transcription, we speed-vacuum concentrated the primer mix into a smaller volume. We then prepared the reaction (0.5 μ M per gene of RT primer (IDT), 1 U μ l⁻¹ RiboLock RNase inhibitor (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 500 μ M dNTPs (Thermo Fisher Scientific), 20 U μ l⁻¹ RevertAid H-Minus M-MuLV reverse transcriptase (Thermo Fisher Scientific) in 1 \times RT buffer). We incubated the samples in the reverse transcription mixture at 37°C overnight. After reverse transcription, we cross-linked the cDNAs in 50 mM BS(PEG)₆ (Thermo Fisher Scientific) for 1 h and neutralized excess cross-linker with 1 M Tris-HCl at pH 8.0 for 30 min, and then washed the sample with PBST twice to eliminate excess Tris buffer. We then prepared a master padlock mix with 200 nM per padlock probe for each target mRNA and speed-vacuum concentrated the mixture for a higher concentration at a smaller volume, if necessary. We ligated the gene padlock probes on the cDNA (200 nM per gene padlock (IDT), 1 U μ l⁻¹ RiboLock RNase Inhibitor, 20% formamide (Thermo Fisher Scientific), 50 mM KCl, 0.4 U μ l⁻¹ RNase H (Qiagen) and 0.5 U μ l⁻¹ Ampligase (Epicentre) in 1 \times Ampligase buffer) for 30 min at 37°C and 45 min at 45°C. Finally, we performed rolling circle amplification (RCA; 125 μ M amino-allyl dUTP (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 250 μ M dNTPs, 5% glycerol and 1 U μ l⁻¹ ϕ 29 DNA polymerase (Thermo Fisher Scientific) in 1 \times ϕ 29 DNA polymerase buffer) overnight at room temperature. After RCA, we again cross-linked the colonies in 50 mM BS(PEG)₆ for 1 h, neutralized with 1 M Tris-HCl at pH 8.0 for 30 min and washed with PBST. We washed the sample in hybridization buffer (10% formamide in 2 \times SSC) and then either added probe detection hybridization solution (0.25 μ M fluorescent probe in hybridization buffer) or gene sequencing primer hybridization solution (1 μ M of sequencing primer in hybridization buffer) for 10 min at room temperature. We then washed the sample with hybridization buffer three times at 2 min each, rinsed the sample in PBST twice, and proceeded to imaging or continued with Illumina sequencing.

BARseq2 simultaneous detection of endogenous genes and barcodes. We prepared a master mix of reverse transcription primers at 0.5 μ M each for all target mRNAs. For volumes exceeding the amount required for reverse transcription, we speed-vacuum concentrated the primer mix into a smaller volume. We then prepared the reaction (0.5 μ M per gene RT primer (IDT), 1 μ M barcode LNA RT primer (Qiagen), 1 U μ l⁻¹ RiboLock RNase inhibitor (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 500 μ M dNTPs (Thermo Fisher Scientific), 20 U μ l⁻¹ RevertAid H-Minus M-MuLV reverse transcriptase (Thermo Fisher Scientific) in 1 \times RT buffer), adding the barcode LNA primer last into the reaction mix to reduce cross-hybridization due to the LNA strong binding affinity. We incubated the samples in the reverse transcription mixture at 37°C overnight. After reverse

transcription, we cross-linked the cDNAs in 50 mM BS(PEG)₆ (Thermo Fisher Scientific) for 1 h and neutralized excess cross-linker with 1 M Tris-HCl at pH 8.0 for 30 min, and then washed the sample with PBST twice to eliminate excess Tris buffer. We then prepared a master padlock mix with 200 nM per padlock probe for each target mRNA and speed-vacuum concentrated the mixture for a higher concentration at a smaller volume, if necessary. We ligated the gene padlock probes on the cDNA (200 nM per gene padlock (IDT), 1 U μ l⁻¹ RiboLock RNase Inhibitor, 20% formamide (Thermo Fisher Scientific), 50 mM KCl, 0.4 U μ l⁻¹ RNase H (Qiagen) and 0.5 U μ l⁻¹ Ampligase (Epicentre) in 1 \times Ampligase buffer) for 30 min at 37°C and 45 min at 45°C. After ligating padlock probes for our target genes, we ligated the padlock probe for the barcode cDNA (100 nM barcode padlock (IDT), 50 μ M dNTPs, 5% glycerol, 1 U μ l⁻¹ RiboLock RNase Inhibitor, 20% formamide (Thermo Fisher Scientific), 50 mM KCl, 0.4 U μ l⁻¹ RNase H (Qiagen), 0.001 U μ l⁻¹ Phusion DNA polymerase (NEB) and 0.5 U μ l⁻¹ Ampligase (Epicentre) in 1 \times Ampligase buffer) without any wash in between, and incubated the reaction for 5 min at 37°C and 40 min at 45°C. We then washed the sample twice with PBST and once with hybridization buffer (10% formamide in 2 \times SSC), before hybridizing 1 μ M of RCA primer in hybridization buffer for 15 min at room temperature. We washed the sample with hybridization buffer three times at 2 min each. Finally, we performed RCA (125 μ M aadUTP (Thermo Fisher Scientific), 0.2 μ g μ l⁻¹ BSA, 250 μ M dNTPs, 5% glycerol and 1 U μ l⁻¹ ϕ 29 DNA polymerase (Thermo Fisher Scientific) in 1 \times ϕ 29 DNA polymerase buffer) overnight at room temperature. After RCA, we again cross-linked the colonies in 50 mM BS(PEG)₆ for 1 h, neutralized with 1 M Tris-HCl at pH 8.0 for 30 min, and washed with PBST. We washed the sample in hybridization buffer (10% formamide in 2 \times SSC) and then added gene sequencing primer hybridization solution (1 μ M of sequencing primer in hybridization buffer) for 10 min at room temperature. We then washed the sample with hybridization buffer three times at 2 min each, rinsed the sample in PBST twice and proceeded to Illumina sequencing.

In situ sequencing of endogenous genes. To sequence the endogenous genes using Illumina sequencing chemistry, we used the HiSeq Rapid SBS Kit v2 reagents to reduce cost from the original sequencing protocol⁶. For the first cycle, we incubated samples in universal sequencing buffer (USB) at 60°C for 3 min, then washed in PBST, followed by incubation in iodoacetamide (9.3 mg in 2 ml PBST) at 60°C for 3 min. We washed the sample in PBST again, rinsed with USB twice more, and then incubated in incorporation mix (IRM) at 60°C for 3 min. We repeated the IRM step again to ensure the reaction was as close to 100% complete as possible. We then washed the sample in PBST once and then continued to wash in PBST four more times at 60°C for 3 min each time. To reduce bleaching during imaging, we imaged the sample in universal scan mix (USM).

For subsequent cycles, we first washed samples in USB, then incubated in cleavage reagent master mix (CRM) at 60°C for 3 min. We repeated the CRM step to ensure complete reaction and washed out residual CRM twice with cleavage wash mix (CWM). We then washed the sample with USB, and then with PBST, before incubating in iodoacetamide at 60°C for 3 min. We repeated this step again to ensure we blocked as many of the free thiol groups as possible to reduce background. We then continued with IRM and PBST washes as described for the first cycle and imaged after each cycle. We performed four sequencing cycles and seven sequencing cycles in total for our cadherins panel of 23 genes and our motor cell-type markers and cadherins panel of 65 genes, respectively.

To visualize high expressors, we cleaved the fluorophores in the last sequencing cycle and washed the sample with CWM and PBST. We then washed our sample in hybridization buffer and added probe detection solution (0.5 μ M for each probe in hybridization buffer) for four different fluorescent probes detecting *Slc17a7*, *Gad1*, *Slc30a3* and all previously sequenced genes, respectively, for 10 min at room temperature. We washed the sample in the same hybridization buffer three times for 2 min each, washed in PBST, before adding DAPI stain (ACDBio) for 2 min at room temperature. We rinsed in PBST again and finally in USM for imaging.

In situ sequencing of barcodes. After sequencing and hybridizing for endogenous genes as described above, we stripped the sample of all hybridized oligonucleotides and sequenced bases by incubating twice in strip buffer (40% formamide in 2 \times SSC with 0.01% Triton-X) at 60°C for 10 min. We washed with PBST, then washed with hybridization buffer, and then incubated samples in barcode sequencing primer hybridization solution (1 μ M sequencing primer in hybridization buffer) for 10 min at room temperature. We washed with hybridization solution three times for 2 min each, before rinsing twice in PBST. We sequenced barcodes with the same sequencing procedure as described for endogenous genes but for 15 cycles in total. At around cycle 4 or 5, we eliminate the iodoacetamide blocker incubation for the rest of sequencing because iodoacetamide blockage is irreversible, so further incubation in this blocker becomes unnecessary after several cycles.

Target area barcode sequencing. Barcode sequencing in target brain areas was performed by the Cold Spring Harbor Laboratory MAPseq Core following procedures used in a previous study⁶. The target areas were dissected to match two other studies in A1 (ref. ⁹) and in M1 (ref. ³) resulting in 11 and 35 projection targets for neurons in auditory cortex and motor cortex, respectively; these projection targets corresponded to most of the major projection targets based on

bulk tracing²¹. A detailed description of each dissected area and correspondence to the Allen reference atlas are shown in Supplementary Table 2.

Fluorescence in situ hybridization. FISH experiments were performed using RNAscope Fluorescent Multiplex Kit v1 according to the manufacturer's protocols with minor modifications to sample preprocessing. For FISH experiments in comparison to BARseq2 endogenous mRNA detection (Figs. 1f and 2e), the samples were fresh frozen in a isopentane bath as described above. From -80°C storage, the samples were immediately submerged in freshly prepared 4% PFA (Electron Microscopy Sciences) for 15 min at 4°C , then dehydrated in 75%, 85% and 100% ethanol twice for 5 min each. After air-drying, we assembled HybriWell-FL chambers (22 mm \times 22 mm \times 0.25 mm; Grace Bio-Labs) and digested the samples in Protease IV for 30 min at room temperature. We washed the samples in PBST, and then proceeded with probe hybridization and subsequent amplification and visualization steps following the manufacturer's protocol, and mounted the samples with coverslips finally for imaging.

For FISH experiments in samples labeled in retrograde, we first imaged the samples before performing FISH. The samples were then dehydrated in 50%, 75% and 100% ethanol twice for 5 min each. After air-drying the samples, we either assembled HybriWell-FL chambers (22 mm \times 22 mm \times 0.25 mm; Grace Bio-Labs) or drew a barrier around the samples using a ImmEdge hydrophobic barrier pen. The samples were then digested in Protease III for 30 min at 40°C , and washed in nuclease-free H_2O twice. We then proceeded to probe hybridization and subsequent amplification and visualization steps following the manufacturer's protocol, and mounted the samples with coverslips for imaging.

For Fig. 1f, the FISH probes used were Mm-Slc17a7-C1, Mm-Slc30a3-C2 and Mm-Cdh13-C3, visualized with Amp4 A It A. For Fig. 2e, the FISH probes used were Mm-Pcdh19-C1, Mm-Cdh8-C2 and Mm-Pcdh20-C3, visualized with Amp4 A It A. For retrograde labeling experiments in Extended Data Fig. 9a–e, the FISH probes used for the cadherins were Mm-Cdh12-C1 (custom ordered, no. 842531), Mm-Cdh8-C1 or Mm-Pcdh19-C1, in addition to Mm-Slc30a3-C2 and Mm-Slc17a7-C3, visualized with Amp4 A It C.

Imaging. All sequencing experiments were performed on an Olympus IX81 microscope with Crest X-light V2 spinning disk confocal, a Photometrics BSI prime camera and an 89North LDI seven-channel laser bank. Retrograde labeling experiments were imaged on either the same microscope or an LSM 710 laser scanning confocal microscope. Filters and lasers used for imaging are listed in Supplementary Table 3. Images were acquired using Micro-Manager (v1.4.23)⁵² on the spinning disk confocal and Zeiss Zen 2012 SP5 FP2 (v14.0.0.0) on the laser scanning confocal.

For all BARseq2 experiments, we imaged endogenous genes using an Olympus UPLFLN $\times 40$ 0.75-NA air objective and tiled 5×5 or 7×5 with 15% overlap between tiles for all sequencing cycles and the hybridization cycles. For each sequencing cycle, the four sequencing channels (G, T, A and C) and the DIC channel were captured. For hybridization cycles, GFP, RFP, Texas Red, Cy5 and DIC channels were captured. At the last cycle (usually the hybridization cycle for high expressors), we also imaged the DAPI channel.

For barcode sequencing, we imaged the first three cycles using the same imaging settings described above at $\times 40$ magnification. The third sequencing cycle was additionally reimaged at $\times 10$ magnification using an Olympus UPLANAPO $\times 10$ 0.45-NA air objective without tiling. All subsequent barcode sequencing cycles were imaged at $\times 10$ magnification.

On the spinning disk confocal, all $\times 40$ BARseq2 and FISH images were acquired as z-stacks with 1- μm step size and 0.16- μm xy pixel size, and all $\times 10$ images were acquired as z-stacks with 5- μm step size.

On the LSM 710, CTB-labeled samples were first imaged using a Plan-Apochromat $\times 10$ 0.45-NA objective without a coverslip as a z-stack with 7- μm z-step size and 0.7- μm xy pixel size. After FISH, the same samples were imaged using a Plan-Apochromat $\times 20$ 0.8-NA objective as a z-stack with 2- μm step size and 0.35- μm xy pixel size.

Probe design. A detailed description of probe sets used for each experiment and their sequences is provided in Supplementary Table 4.

To design reverse transcription primers and padlock probes, we tried to design as many probe sets as possible on each transcript while avoiding the end (~ 20 nucleotides) of the mRNA transcripts and ensuring at least a 3-nucleotide-long gap between two adjacent probe sets. Specific reverse transcription primers were designed to be 25 to 26 nucleotides long with amino modifier C6 at the 5' end and purified by high-performance liquid chromatography. In addition, we avoided sequences that contained G/C quadruplexes and/or had a low melting temperature, T_m (below 55°C). Padlock probes were designed to have two arms of 21 to 23 nucleotides long with a minimum T_m of 58°C , GC content between 40% and 60% and high complexity. The two arms were connected by a backbone consisting of a 32-nucleotide-long sequencing primer or detection probe target site, a 7-nucleotide gene-specific index, and a 3-nucleotide-long 3' linker. For padlock probes designed for hybridization readout, different backbone sequences were used for different genes. We further filtered out padlock probe sequences with potential nonspecific binding. To find potential nonspecific binding targets, we blasted the ligated

padlock arm sequences against the mouse genome and identified all targets with (1) 3 nucleotides of perfect match on either side of the ligation junction, (2) no gap and/or insertion within 7 nucleotides of the ligation junction and (3) melting temperatures of at least 37°C for nonspecific binding of each arm.

We maximized the number of padlock probe sets for *Slc17a7* (23 probes), *Slc30a3* (19 probes), *Gad1* (24 probes) and *Cdh13* (30 probes). These probe sets were used to evaluate the relationship between detection sensitivity and probe numbers. For the cadherin panels and the cell-type marker panels, we selected a subset of probes for each gene so that we had at most 12 probe sets per gene. Some shorter genes had fewer than 12 probes. These panels resulted in sensitivity that was sufficient for the present experiments, albeit somewhat below the maximum achievable with more probes. All but three genes (*Slc17a7*, *Slc30a3* and *Gad1*) were visualized using combinatorial GII codes (4 nucleotides in auditory cortex and 7 nucleotides in motor cortex; Supplementary Table 4); only a small subset of all possible GIIs were used, ensuring a Hamming distance of at least two bases between all pairs of GIIs in auditory cortex (from 4 nucleotides) and three bases in motor cortex (from 7 nucleotides) for error correction. The three remaining genes with high expression (*Slc17a7*, *Gad1* and *Slc30a3*) were detected by hybridization.

Optimization of endogenous mRNA detection. We optimized padlock probes, tissue pretreatment and reverse transcription to maximize detection sensitivity. We found that using multiple padlocks per mRNA transcript, with each padlock targeting a different site on the mRNA coding sequence, increased detection efficiency substantially (Fig. 1e). The increase in sensitivity varied across genes, but this was likely caused by differences in sensitivity of the single probe to which we normalized the sensitivities. For tissue pretreatment, we found that thin fresh-frozen tissue cryosections fixed with 4% PFA for 30 min to 1 h (Extended Data Fig. 1a) yielded higher mRNA sensitivity than shorter fixation or other pretreatments, such as PFA-perfused tissue slices with or without post-fixation. For reverse transcription, we found that reverse transcription primers specific to the targets at a concentration of 0.5–5 μM each yielded higher sensitivity than using random primers at concentrations up to 50 μM (Extended Data Fig. 1b). Altogether, these optimizations were crucial for increased mRNA detection sensitivity comparable to hybridization-based techniques.

To quantify the sensitivity of BARseq2 compared to conventional FISH methods, we detected two genes, *Slc30a3* and *Cdh13*, using both BARseq2 and RNAscope (Fig. 1f). We also probed for a third gene, *Slc17a7*, but at the resolution we imaged at, we were unable to fully resolve the signals from both BARseq2 and RNAscope; therefore, we only used *Slc30a3* and *Cdh13*, not *Slc17a7*, to evaluate the sensitivity of BARseq2. Linear regression between BARseq2 and RNAscope counts of *Slc30a3* and *Cdh13* genes in these two genes resulted in a slope of 1.65 (Extended Data Fig. 1c,d; $R^2 = 0.73$), indicating that BARseq2 achieved a sensitivity of about $1/1.65 \approx 60\%$, compared to RNAscope.

To multiplex gene detection with high imaging throughput, we optimized in situ sequencing to robustly read out GIIs of single rolonies over many sequencing cycles. We had previously adapted Illumina sequencing chemistry to sequence neuronal somata filled abundantly with RNA barcode rolonies, that is, DNA nanoballs generated by RCA^{6,35}. However, directly applying this method to sequence single rolonies generated from individual mRNAs proved difficult due to heating cycles and harsh stripping treatments that led to loss and/or jittering of rolonies (Extended Data Fig. 1e). To allow robust sequencing of single rolonies, we optimized cryosectioning and amino-allyl dUTP concentration³³ to cross-link rolonies more extensively, achieving less spatial jitter of single rolonies between imaging cycles (Extended Data Fig. 1e–h) and stronger signals (Extended Data Fig. 1i,j) retained over cycles. This robust in situ sequencing of combinatorial GII codes allowed BARseq2 to achieve fast imaging critical for high-throughput correlation of gene expression with projections.

Simultaneous detection of endogenous mRNAs and barcodes using BARseq2.

To assess multiplex gene expression and long-range projections in the same cells, we optimized for simultaneous detection and amplification of both endogenous mRNAs and barcodes. Although both endogenous mRNAs and barcodes are amplified using padlock probe-based approaches, amplifying barcodes required the addition of a DNA polymerase to copy barcode sequences into padlock probes to allow direct sequencing of diverse barcodes (up to $\sim 10^{18}$ diversity; Fig. 1c). Directly combining the two processes reduced the detection sensitivity of target mRNAs due to the addition of the DNA polymerase (Extended Data Fig. 6a; $37\% \pm 3\%$ (mean \pm s.d.); comparing the control condition to the zero-polymerase concentration). To preserve detection sensitivity for endogenous mRNAs while allowing the sequencing of diverse barcodes, we adjusted the concentration of the DNA polymerase to 0.001 $\text{U } \mu\text{l}^{-1}$ (1/200 of the amount in the original BARseq), which doubled the sensitivity for endogenous mRNAs while also maintaining the sensitivity for barcodes (Extended Data Fig. 6a). This optimization allowed BARseq2 to detect both endogenous mRNAs and RNA barcodes together in the same neurons without compromising sensitivity.

Single-cell RNA-seq of auditory cortex. To dissociate neurons for scRNA-seq, we anesthetized animals with isoflurane and decapitated the animals. We then used a 2-mm biopsy punch to remove the auditory cortex. The tissue was then

dissected in ice-cold HABG medium (40 ml Hibernate A (Brainbits), 0.8 ml B27 (Thermo Fisher Scientific) and 0.1 ml Glutamax (Thermo Fisher Scientific)) into small pieces and digested in 3 ml prewarmed papain solution (3 ml Hibernate A-Ca (Brainbits), 6 mg papain (Brainbits) and 7.5 μ l Glutamax) at 30 °C for 40 min. The digested tissues were then triturated in 2 ml prewarmed HABG for ten times using a salinized pipette with a 500- μ m opening. The undissociated tissues were transferred to a new tube with 2 ml HABG and triturated another ten times. The undissociated tissues were transferred again to a new tube with 2 ml HABG and triturated five times. The three tubes of HABG were combined and laid on top of a density gradient of 17.3%, 12.4%, 9.9% and 7.4% (vol/vol) Optiprep (Sigma) in HABG and centrifuged at 750g for 15 min. After removing the top two fractions, we collected the next two and half fractions and diluted in 5 ml HABG and centrifuged at 300g for 5 min. The pellet was washed in 5 ml HABG, pelleted again and resuspended in 100 μ l HABG. The cell suspension was then processed for library preparation using 10x Genomics Chromium Single Cell 3' Kits v3 according to the manufacturer's protocol. One of the scRNA-seq datasets was previously published⁵, and a new dataset was obtained in this study.

BARseq2 data processing. Sequencing data for projection target areas were acquired through the MAPseq core facility at Cold Spring Harbor Laboratory. We first demultiplexed raw sequencing reads and applied a threshold by read counts per molecule to remove PCR errors. This produced a list of unique barcode sequences with molecule counts in each target area. We then corrected for sequencing and amplification errors, allowing up to three mismatches. The resulting error-corrected barcode molecule counts were used to generate the projection matrix.

To process in situ sequencing data for genes, we first performed maximum projection of the image stacks along the z axis. Each maximum projection image was then corrected for sequencing channel bleed-through and lateral shift across channels. The images were then filtered with a median filter and background subtracted using a rolling ball with a radius of ten pixels. The sequencing cycle images were then registered to the first sequencing cycle using the sum of all four sequencing channels, and the hybridization images were registered to the first sequencing cycle using the channel that labeled all sequenced colonies. Registrations were performed by maximizing enhanced cross-correlation⁵⁴. After all images were registered, putative colonies were then picked from the first sequencing cycle by finding all peaks that were at least brighter than all surrounding valleys by a certain threshold determined empirically. This was achieved by first performing morphological reconstruction using the original image as the mask and the image minus the threshold as the marker, followed by identification of all local maxima. We then deconvolved all registered images and found the signal intensities for all colonies across all sequencing cycles and channels.

At this point, the signal for each colony is represented by an $m \times 1$ vector, in which m equals four (sequencing channels) times the number of cycles. To identify the gene that each colony corresponds to, we project the signal vector onto the signal vector of all genes and find the two genes with the highest projections, I_1 and I_2 . For colonies whose $(I_1 - I_2)/I_1$ is above a threshold, we assign the genes with the highest projections to these colonies. The remaining colonies are filtered out. For hybridization cycles, the channel in which the colonies are found is used directly to identify the genes.

For experiments in which genes were detected without barcodes for projection mapping, we segmented somas based on the colony signals, background fluorescence from somas and nuclear staining using Cellpose⁵⁵, and assigned the colonies to the segmented cells.

For experiments in which genes were detected in conjugation with barcodes, we further registered barcode sequencing cycles to the first sequencing cycle for genes using the DIC channel. The barcode sequencing images were then filtered with a median filter and background subtracted using a rolling ball with a radius of 50 pixels. The high-resolution images for the second and third cycles were then registered to the first sequencing cycle of barcodes using the sum of all four sequencing channels. The low-resolution images of the third sequencing cycle were then registered to the high-resolution image of the same cycle.

To segment the barcoded cells from the high-resolution images, we first determined 'seed' pixels by identifying local maxima in the first sequencing cycle image as described above. These seed pixels are positions of the strongest signal within putative cell bodies. Then for each seed pixel, we calculated the projection of signal vectors for all other pixels within a local area on the signal vector of the seed pixel and the rejection of signal vectors for these pixels from the signal vector of the seed pixel. We then segmented the cell bodies by finding all pixels that fulfill the following criteria: (1) the projections of their signal vectors are above a threshold; (2) the ratios between the rejections and projections are below a threshold; and (3) they are connected to the seed pixel. In parallel, we performed a second segmentation using only the DAPI signals and gene sequencing images with a marker-based watershed without using the barcode sequencing images, and found the segmented cells that overlapped with the barcode segmented cells. We then visually inspected the sequencing images and segmentations for each cell to determine which segmentation produced better results and to eliminate badly segmented cells. We then assign gene colonies to the filtered segmented cells to produce the expression matrix.

To find the barcode sequences of the segmented cell, we integrated signals over the whole segmented cells and called the channel with the strongest signal as the base in both the high-resolution images and the low-resolution images. We then concatenated the sequences from the high-resolution images and the low-resolution images to produce the full barcode sequences. To find the projection patterns, these in situ sequenced barcodes were then matched to the barcodes identified in the projection areas allowing one mismatch but not ambiguous matches (that is, one in situ barcode matching to multiple barcodes found in projection sites).

Analysis of BARseq2 gene expression data. All analyses were carried out in MATLAB. For analysis of gene-only datasets, neurons were first filtered by requiring at least ten counts of *Slc17a7* or *Gad1* and were positioned within the cortex. To make the data comparable to previous studies⁶, the cortical depths of neurons were normalized to a total thickness of 1,200 μ m for auditory cortex and 1,500 μ m for motor cortex. To find cadherins that were differentially expressed in cell types, the expression of cadherins in each cell type was compared to the expression of cadherins in all other cell types using rank-sum tests.

Laminar distribution of cadherins. Because many genes, especially cell adhesion molecules, are differentially expressed across cortical layers, we evaluated how well BARseq2 can capture spatial organization of cadherins compared to existing methods, such as FISH. To compare laminar distribution observed by BARseq2, FISH and Allen Brain Atlas, we quantified gene expression signal densities across 100- μ m bins in laminar depth. For BARseq2 and FISH, the quantification was performed by counting dots. For Allen Brain Atlas, the quantifications were done by integrating signal intensities over all pixels in each bin. Because each bin had a different number of pixels sampled in our data, we then divided the gene expression signals by the area observed in the images to calculate the density. We then z -scored the densities within each gene to produce the laminar profiles for each gene.

RNAscope against *Cdh8*, *Pcdh19* and *Pcdh20* revealed laminar expression profiles that were qualitatively similar to those obtained by BARseq2 (Fig. 2e). For *Pcdh20*, the dynamic range of gene expression (that is, the differences between peaks and valleys in expression) was more pronounced in the BARseq2 data than that observed by RNAscope. Because low sensitivity and/or low specificity would likely result in a reduction, not an increase, in the dynamic range of expression, it is unlikely that such quantitative differences in the laminar profiles of gene expression were caused by sensitivity and/or specificity issues with BARseq2. We suspect that the reduced dynamic range in RNAscope is caused by nonspecific signals inherent to amplified FISH methods. We therefore sought to compare BARseq2 to other FISH datasets to confirm its accuracy.

We then compared the distributions of genes obtained by BARseq2 to those in the Allen gene expression atlas²¹ (Fig. 2f and Extended Data Fig. 3). The laminar distribution of gene expression revealed by BARseq2 was highly correlated with that in the Allen gene expression atlas (Spearman correlation $\rho = 0.696$, $P = 3.8 \times 10^{-29}$). Specifically, the laminar distribution of *Pcdh20* obtained by BARseq2 matched very well with *Pcdh20* in the Allen gene expression atlas (Extended Data Fig. 3). These results indicate that BARseq2 accurately captured the laminar distribution of cadherin expression.

Gene-pair expression in single neurons. To test whether BARseq2 accurately captures gene expression, we compared the expression of two pairs of genes in single neurons. First, we compared the expression of *Slc17a7* and *Gad1*, two genes that are expressed in two distinct classes of neurons. Second, we compared the expression of *Slc30a3* and *Cdh24*, two genes that are anti-correlated at the subtype level based on scRNA-seq³.

Slc17a7 and *Gad1* are expressed in excitatory and inhibitory neurons, respectively. They are thus almost never expressed in the same neuron in the cortex. To quantify the mutual exclusivity of *Slc17a7* and *Gad1* in neurons, we defined the exclusivity index $E = P(Gad1|Slc17a7)/P(Gad1)$, where $P(Gad1|Slc17a7)$ indicates the probability of a cell expressing at least ten counts of *Gad1* conditioned on the expression of at least ten counts of *Slc17a7*, and $P(Gad1)$ indicates the probability of a cell expressing at least ten counts of *Gad1* in all filtered neurons.

BARseq2 recapitulated the mutual exclusivity between these two genes (Fig. 2j,k), but a small number of neurons did express both *Slc17a7* and *Gad1* (gray cells in Fig. 2j). This could be caused by overlapping cells (that is, an inhibitory neuron and an excitatory neuron at the same x/y position, but in different z planes were merged together in the maximum projection images) or cell segmentation errors (two adjacent cells incorrectly segmented as a single cell). Because the sections we used were 10- μ m thick, comparable to the diameter of an average neuron, the latter source of error was likely to be more common.

This type of error was similar to doublets in droplet-based scRNA-seq techniques. Assuming that the mutual exclusions of *Slc17a7* and *Gad1* were absolute, then we could estimate the 'doublet' rate as the ratio between the probability of neurons expressing both genes and the product of the probabilities of neurons expressing either gene. Using this formula, we estimated the doublet rate of BARseq2 to be 7.5%, which is in a similar range as droplet-based scRNA-seq

techniques (usually <5%). Improvement in cell segmentation algorithms may further reduce the doublet rate.

In addition to cells that express both *Gad1* and *Slc17a7* at substantial levels, most cells that expressed one of the two genes dominantly also had nonzero expression of the other gene, albeit at much lower levels. This noise floor could be caused by mRNAs in dendrites that were incorrectly assigned to other neurons. Because the expression levels of these genes in the somata were much higher than those in the dendrites, this type of error was unlikely to substantially affect the determination of excitatory and inhibitory neurons.

Similarly, consistent with a previous scRNA-seq study³, BARseq2 also confirmed the observation that *Slc30a3* was more highly expressed in subtypes of excitatory neurons that did not express *Cdh24* compared to projection neurons that did express *Cdh24* (Extended Data 5a,b; $P = 5 \times 10^{-26}$ using two-tailed rank-sum test on scRNA-seq data using Smart-Seq2 ($n = 10,044$ neurons)³, and $P = 4 \times 10^{-65}$ on BARseq2 data ($n = 2,947$ neurons)).

Cell typing in BARseq2 and single-cell data. To select a panel of marker genes, we chose meta-analytic markers from seven scRNA-seq datasets in the motor cortex²³, accessed from the NeMO archive. In each dataset and for each cell type, we extracted differentially expressed genes (DEGs) among excitatory neurons ('glutamatergic' class, one-versus-all DEGs, fold change > 2, Mann-Whitney FDR < 0.05). We filtered out genes with low expression (average counts per million (CPM) < 100), then ranked genes primarily by the number of datasets where they were DEGs and secondarily by average fold change, and selected the top five markers.

To examine if multiplexing affects detection sensitivity, we probed for *Slc17a7*, *Slc30a3* and *Gad1* either as a separate three-gene panel or as part of the 65-gene panel (20 cadherins and 45 marker genes). The mean expression densities across laminar positions for the three genes were similar between the three-gene panel and the 65-gene panel (Extended Data Fig. 5c; $P = 0.22$ for *Slc17a7*, $P = 0.49$ for *Slc30a3* and $P = 0.66$ for *Gad1* using two-tailed rank-sum tests), suggesting that targeting more genes did not affect detection sensitivity of each gene.

To call cell types in BARseq2 and single-cell data, we used the following procedure. First, we normalized counts to $\log(1 + \text{CPM})$, then we computed the average marker expression for each cell type and assigned the cell type with the highest average expression. If two marker sets were tied for highest expression, the cell was left unassigned. This method of cell typing achieved good precision and recall for most cell types when applied to scRNA-seq data (Extended Data Fig. 5d). We applied the procedure across nine datasets to check whether it is robust across technologies and sequencing depth (Extended Data Fig. 5e,f). Overall, we observed extremely high performance for NP and CT subtypes in all cases, while L6b was slightly better predicted in high-depth datasets. The cell-typing method always predicted IT cells correctly, but not always the correct layer (L2/3, L5, L6 and Car3; Extended Data Fig. 5g). This is consistent with the observation that IT types form a continuum in single-cell datasets, making it difficult to fully separate subtypes by layer. Finally, the PT type proved to be the most difficult cell type to predict. While all PT cells were correctly annotated as PT (Extended Data Fig. 5h), numerous L2/3 IT and L5 IT cells were wrongly annotated as PT, in particular in high-depth datasets (Extended Data Fig. 5f,g). We believe that this was due to an imbalance in the marker panel, with PT markers showing higher expression than markers from other cell types. We tested various normalization procedures to overcome this effect but found that results were insensitive to normalization overall (Extended Data Fig. 5f).

Using this panel and cell-typing method, we determined the transcriptomic types of excitatory neurons in motor cortex using BARseq2 (Fig. 3b). Most transcriptomic types were found enriched in the correct layers. One exception to this was the L6 Car3 IT type. In general, few L6 Car3 IT neurons were identified by BARseq2. Furthermore, even though L6 Car3 IT neurons were predominantly in L6, some were identified in L2/3 by BARseq2 (Fig. 3c). This result was surprising, given that L6 Car3 IT neurons, when present, were only rarely mistyped as L2/3 in our preliminary analyses (Extended Data Fig. 5g). L6 Car3 IT neurons were only rarely detected in the datasets used to select markers, so we expect that using additional data will lead to a more robust marker selection and better cell-typing performance with BARseq2. These optimizations, however, are beyond the scope of this paper.

Gene expression in barcoded neurons. Gene expression in Sindbis-infected barcoded neurons largely reflect the gene expression in non-barcoded neurons. For example, the expression of the excitatory marker *Slc17a7* and the inhibitory marker *Gad1* remained mutually exclusive in barcoded neurons in both auditory cortex and motor cortex (Extended Data Fig. 6c,d). This mutual exclusivity was preserved despite an overall reduction in mRNA expression (Extended Data Fig. 6e; median read of 38 in barcoded cells in both auditory and motor cortex, compared to 64 and 48 in non-barcoded cells in the two cortical areas, respectively). Similarly, *Slc30a3* remained differentially expressed across barcoded excitatory neurons with or without *Cdh24* expression as it was in non-barcoded excitatory neurons (Extended Data Fig. 6f; $P = 1 \times 10^{-6}$ using rank-sum test, $n = 810$ neurons). Although our observations cannot rule out the possibility that a small subset of genes (for example, viral response genes) may be disrupted by Sindbis infection, these results

suggest that the coexpression relationships of most genes in Sindbis-infected neurons reflect those in noninfected cells.

Analysis of BARseq2 gene expression and projection dataset. For analysis of BARseq2 datasets with both gene expression and projections, we first evaluated the mutual exclusivity of *Slc17a7* and *Gad1* expression (see below). For this purpose, the neurons were filtered with the same thresholds as in the gene-only dataset. For all other analyses, we used a more relaxed filtering to compensate for the reduced gene expression in barcoded cells, requiring neurons to have at least five counts of *Slc17a7* or *Gad1*. In this filtered set, neurons were considered excitatory if the counts of *Slc17a7* were larger than the counts of *Gad1*, and were considered inhibitory if the counts of *Gad1* were larger than the counts of *Slc17a7*. Projection data were log normalized as in previous studies⁶. We further normalized the projection strengths of each area to two previous clustered BARseq datasets⁶ and used a random forest classifier to assign neurons to projection clusters.

To find cadherins that were differentially expressed across major projection classes and between auditory and motor cortex, we performed rank-sum tests for pairwise comparisons among major classes or the two areas for each cadherin and calculated the FDRs.

Projection modules were identified using NMF³². To find the variance in projections explained by cadherins and/or laminar positions (Extended Data Fig. 8), we used Gaussian process regression to predict projection modules using the laminar position of neurons as a predictor and linear regression to predict projection modules using the expression of individual cadherins. The variance explained by each predictor was reported after 100 iterations of ten-fold cross validation. To find cadherins that were associated with projection modules, we calculated the Spearman correlation between the coefficients for projection modules and gene counts. To generate the plots of differential gene expression in Fig. 6e, we sorted the neurons by the coefficients for projection modules and smoothed gene expression using a window of 101 neurons.

Projections of excitatory and inhibitory neurons. BARseq2 accurately observed the fact that projection neurons in the cortex are predominantly excitatory and express the excitatory marker *Slc17a7*, not the inhibitory marker *Gad1*. To distinguish between excitatory and inhibitory neurons, we categorized a neuron as excitatory or inhibitory if (1) the neuron had higher expression of the excitatory marker *Slc17a7* or the inhibitory marker *Gad1*, respectively, and (2) the marker was expressed at greater than five reads in the cell. This threshold resulted in 2,496 excitatory neurons (947 in auditory cortex and 1,549 in motor cortex) and 240 inhibitory neurons (100 in auditory cortex and 140 in motor cortex; Fig. 4d). Consistent with previous observations, most cortical projection neurons identified by BARseq2 were excitatory (Fig. 4e). However, we also identified a small fraction of inhibitory projection neurons. Some of these neurons could be caused by 'doublets' as discussed above. Consistent with this hypothesis, the inhibitory projection neurons (and some excitatory projection neurons) in motor cortex expressed both *Gad1* and *Slc17a7* at similar levels (Extended Data Fig. 6g). However, inhibitory projection neurons in auditory cortex expressed only *Gad1*, not *Slc17a7* (Extended Data Fig. 6h), suggesting that these were real inhibitory projection neurons. This observation was consistent with previous reports of rare inhibitory projection neurons in the cortex^{6,56}. We did not further analyze these inhibitory projection neurons.

We also observed many excitatory neurons without projections (Fig. 4d,e), similar to those observed in previous BARseq experiments⁶. These neurons were likely non-projecting excitatory neurons and neurons that project only locally or to neighboring cortical areas³ that we did not sample.

Differential expression of cadherins across IT, PT and CT neurons. BARseq2 revealed differential gene expression across major classes of neurons defined by projections. We found that many cadherins (8 for auditory cortex and 12 for motor cortex) were differentially expressed across IT, PT and CT neurons that were defined by projections as in previous studies^{6,6} (Fig. 5a–c). Several cadherins were consistently differentially expressed in both cortical areas. For example, *Cdh6* and *Cdh13* were overexpressed in PT neurons compared to the other two classes, whereas *Cdh8* was underexpressed in CT neurons compared to the other two classes (FDR < 0.05, rank-sum test). In addition, we also found nine cadherins that were differentially expressed across the two cortical areas in at least one class (Fig. 5d; FDR < 0.05, rank-sum test).

Major classes of projection neurons (IT, PT and CT) differ in both gene expression and projection patterns. Therefore, the differential expression of cadherins observed across these three major classes defined by projection patterns should be consistent with the differential expression across the classes defined by transcriptomic methods. To test this, we compared the differences in mean expression of cadherins in the three classes in motor cortex and auditory cortex observed by BARseq2 to those observed using scRNA-seq in neighboring cortical areas (V1 and ALM)³. Generally, differentially expressed cadherins identified by BARseq were also differentially expressed in scRNA-seq (Extended Data Fig. 7a; the rank correlation of the differences in cadherin expression across major neuronal types was 0.61 between BARseq and scRNA-seq, compared to 0.39 between auditory cortex and motor cortex in BARseq). Importantly, all cadherins that were consistently differentially expressed in both A1 and M1 were also differentially

expressed across the same pairs of major classes in V1 and ALM as shown by scRNA-seq (Extended Data Fig. 7a). Several cadherins, including *Pcdh7* and *Cdh11*, were differentially expressed with the opposite signs in scRNA-seq and in BARseq2 (Extended Data Fig. 7a). However, these cadherins were not consistently expressed across motor and auditory cortex. For example, *Pcdh7* was expressed at a significantly higher level in PT neurons than CT neurons in motor cortex ($P < 10^{-8}$; Fig. 5c), but at a lower level in PT neurons than CT neurons in auditory cortex ($P = 0.0011$, not statistically significant at $FDR < 0.05$). It is thus likely that these differences between observations by BARseq2 and by scRNA-seq reflect area-to-area differences, not methodological differences. These results confirm the differential expression of cadherins across major classes identified by BARseq2.

Projection differences across transcriptionally defined IT subtypes. BARseq2 confirmed known biases in projection patterns across transcriptomic IT subtypes in auditory cortex (Extended Data Fig. 7b,c). Previous studies using both barcoding-based strategies and single-cell tracing have identified distinctive projection patterns for two transcriptomic subtypes of IT neurons, IT3 (L6 IT) and IT4 (L6 Car3 IT)^{6,26}. To test if we could capture the same projection specificity of transcriptomic subtypes, we mapped projection patterns to projection clusters identified in a previous study in auditory cortex, and used a combination of gene expression and laminar position to distinguish four transcriptomic subtypes of IT neurons⁶. These subtypes were defined consistently with a previous study⁶ for ease of comparison. Specifically, we defined IT1 as neurons with depths of less than 590 μm , IT2 as neurons with depths between 590 and 830 μm and did not express *Cdh13*; IT3 as neurons between 590 and 830 μm that expressed *Cdh13* or neurons deeper than 830 μm that expressed *Slc30a3*; and IT4 as neurons deeper than 830 μm that did not express *Slc30a3*.

As expected, the two transcriptomic subtypes (IT3 and IT4) predominantly found in L5 and L6 were indeed more likely to project only to the ipsilateral cortex, without projections to the contralateral cortex or the striatum ($P = 4 \times 10^{-7}$ comparing the fraction of neurons with only ipsilateral cortical projections in IT3/IT4 to the fraction of them in IT1/IT2 using Fisher's test; Extended Data Fig. 7b,c). Between IT3 and IT4, IT4 neurons were more likely to project ipsilaterally (58% of IT3 neurons compared to 92% of IT4 neurons; $P = 1 \times 10^{-4}$ using Fisher's test), whereas IT3 neurons were more likely to project contralaterally (66% of IT3 neurons compared to 14% of IT4 neurons, $P = 5 \times 10^{-8}$ using Fisher's test). Thus, BARseq2 recapitulated known projection differences across transcriptomic subtypes of IT neurons.

Cadherin coexpression module analysis. To extract robust modules of coexpressed cadherins, we used a previously developed approach to combine multiple datasets by meta-analysis, a crucial step to attenuate technical and biological noise^{33,34}. Briefly, we built coexpression networks using the Spearman correlation for seven scRNA-seq datasets in the motor cortex²³, accessed from the NeMO archive and subset to the following subclasses: 'L2/3 IT', 'L4/5IT', 'L5 IT', 'L6 IT' and 'L6 IT Car3'. We ranked each network, then averaged the networks to obtain our final meta-analytic network. We then applied hierarchical clustering with average linkage and extracted modules using the dynamic tree-cutting tree algorithm³¹.

To compute the association between coexpression modules and projection patterns, we framed the association as a classification task: can we predict projection patterns from module expression? First, we generated labels by binarizing each projection pattern—cells with a projection strictly greater than the median projection strength were marked as positives. Next, we generated predictors by computing gene module expression as the average $\log(\text{CPM} + 1)$ across all genes in the module. We reported the association strength (classification results) as an AUROC. To compute the association between coexpression modules and cell types, we used a similar approach, using clusters defined by the BRAIN Initiative Cell Census Network²³ as labels. For visualization, cell types were organized according to the following procedure: cell types were reduced to a centroid by taking the median expression for each gene, then cell types were clustered according to hierarchical clustering with average linkage with correlation-based distance.

Validation of cadherin correlates of IT projections using in situ hybridization and retrograde labeling. To confirm that *Cdh8*, *Cdh12* and *Pcdh19* correlated with ipsilateral, contralateral and striatal projections, respectively, we performed CTB retrograde labeling from the projection targets and performed FISH against *Slc17a7*, *Slc30a3* and the cadherins in both A1 and M1 (Extended Data Fig. 9a; see Supplementary Table 1 for injection coordinates). We then quantified cadherin expression and CTB labeling in IT neurons that had sufficient DAPI signals and expressed both *Slc17a7*, an excitatory cell marker, and *Slc30a3*, which labeled the majority of IT neurons (Extended Data Fig. 9b). Neurons that had weak and/or ambiguous CTB signals were excluded from the analyses. Indeed, we observed that the three cadherins were expressed at higher levels in CTB⁺ neurons in both areas despite notable overlap in expression between CTB⁺ and CTB⁻ neurons (Extended Data Fig. 9c–e). This overlap was expected because CTB was unlikely to have labeled all neurons that projected to the areas that we sampled with BARseq2. For example, in a previous study, we found that less than half of neurons with

projections detected by BARseq were also labeled by injection of CTB into the same target area⁶. These results thus provide further support for the finding that cadherins correlate with similar projections in both A1 and M1.

Statistics and reproducibility. No statistical method was used to predetermine sample size, but our sample sizes are similar to those reported in previous publications^{6,14}. No data were excluded from the analyses. Because only wild-type animals were used and the findings did not rely on comparison across animals, the experiments were not randomized and the investigators were not blinded to allocation of animals during experiments and outcome assessment. All statistical tests performed are indicated in the text. Two-tailed tests and Bonferroni correction were used for all P values reported unless noted otherwise. Wherever indicated, FDRs were computed according to the Benjamini–Hochberg procedure⁵⁷. All statistical tests used were non-parametric except when statistical significance was estimated for the Pearson correlation (Fig. 6a). When estimating statistical significance for the Pearson correlation, normal distribution was assumed, but this was not formally tested.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw target area sequencing data (Fig. 4c; SRR12247894, SRR12245390 and SRR12245389) and scRNA-seq data (Fig. 2g–i) are deposited at the Sequence Read Archive (SRR13716225). Raw in situ sequencing images (Figs. 2–4) are deposited at the Brain Image Library (<https://download.brainimaginglibrary.org/06/35/0635a0b3b0954c7e/>). Example annotated images from the dissected brain slices and other data and intermediate processed sequencing data are deposited at Mendeley Data (<https://doi.org/10.17632/jnx89bmv4s.2>).

Code availability

Processing scripts are deposited at Mendeley Data (<https://doi.org/10.17632/jnx89bmv4s.2>).

References

- Oh, S. W. et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
- Edelstein, A. D. et al. Advanced methods of microscope control using μ Manager software. *J. Biol. Methods* **1**, e10 (2014).
- Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
- Evangelidis, G. D. & Psarakis, E. Z. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 1858–1865 (2008).
- Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).
- Rock, C., Zurita, H., Wilson, C. & Apicella, A. J. An inhibitory corticostriatal pathway. *Life* **5**, e15890 (2016).
- Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

Acknowledgements

The authors thank members of the MAPseq core facility, H. Zhan, Y. Li and N. Gemmill, for MAPseq data production; K. Matho and Z. J. Huang for dissection coordinates in motor cortex; H. Zhan, L. Yuan, H. L. Gilbert, K. Matho, J. Kecsich and D. Fürth for useful discussions; and W. Wadolowski, B. Burbach, K. Lucere and E. Fong for technical support. This work was supported by the National Institutes of Health (5RO1NS073129, 5RO1DA036913, RF1MH114132 and U01MH109113 to A.M.Z.; R01MH113005 and R01LM012736 to J.G.; and U19MH114821 to A.M.Z. and J.G.), the Brain Research Foundation (BRF-SIA-2014-03 to A.M.Z.), IARPA MICrONS (D16PC0008 to A.M.Z.), Paul Allen Distinguished Investigator Award (to A.M.Z.), Simons Foundation (350789 to X.C.), Chan Zuckerberg Initiative (2017-0530 ZADOR/ALLEN INST (SVCF) SUB to A.M.Z.) and Robert Lourie award (to A.M.Z.). This work was additionally supported by the Assistant Secretary of Defense for Health Affairs endorsed by the Department of Defense, through the FY18 PRMRP Discovery Award Program (W81XWH1910083 to X.C.) Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the U.S. Army. In conducting research using animals, the investigators adhered to the laws of the United States and regulations of the Department of Agriculture.

Author contributions

Y.-C.S., X.C. and A.M.Z. conceived the study. Y.-C.S. and X.C. optimized and performed BARseq2. Y.-C.S., X.C. and H.Z. collected BARseq2 data. X.C., S.F. and J.G. analyzed data. Y.-C.S., X.C. and S.F. selected gene panels. X.C. and S.L. compared gene expression between BARseq2 and Allen ISH. Y.-C.S. and X.C. performed retrograde tracing combined with FISH validations. Y.-C.S., X.C., S.F. and A.M.Z. wrote the paper.

Competing interests

A.M.Z. is a founder and equity owner of Cajal Neuroscience and a member of its scientific advisory board. The remaining authors declare no competing interests.

Additional information

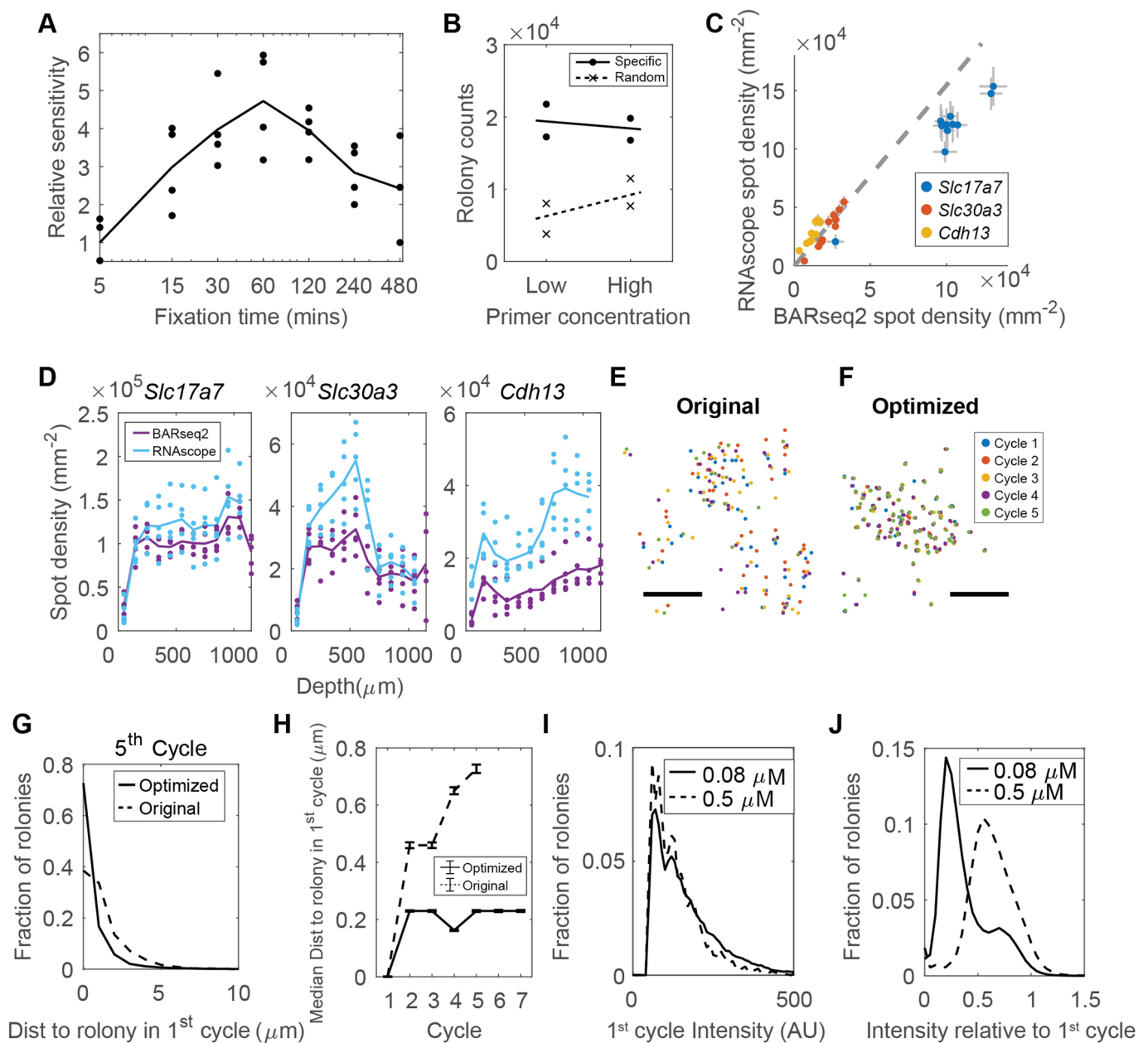
Extended data is available for this paper at <https://doi.org/10.1038/s41593-021-00842-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41593-021-00842-4>.

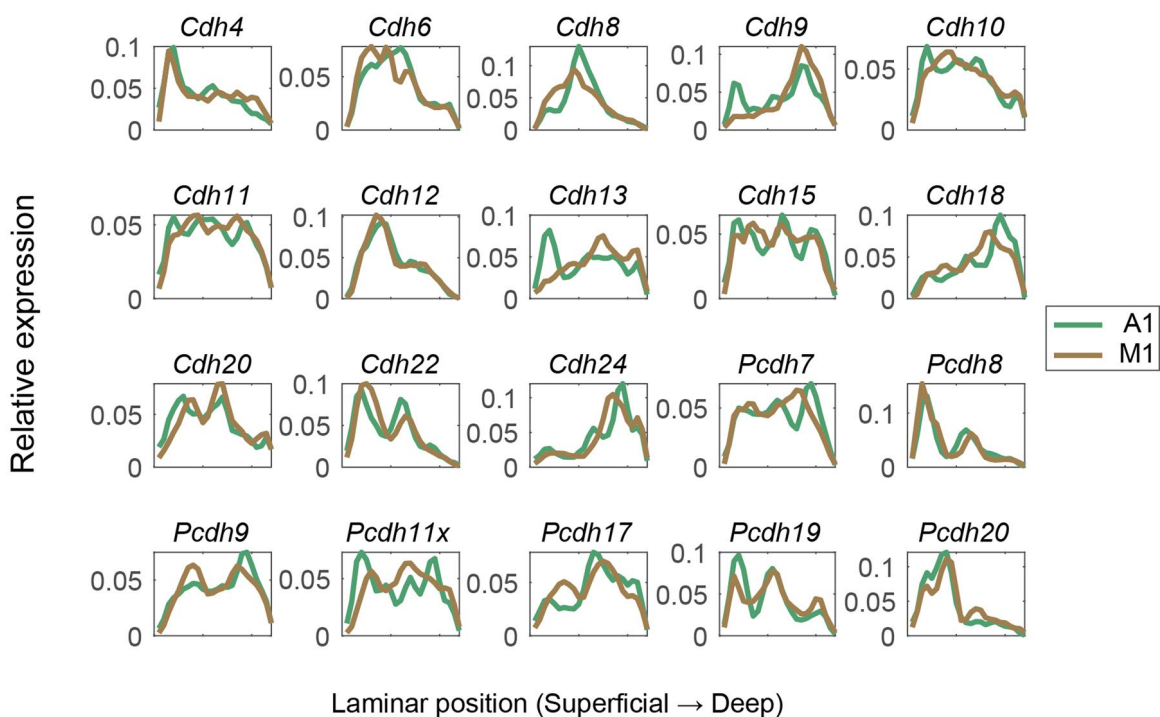
Correspondence and requests for materials should be addressed to X.C. or A.M.Z.

Peer review information *Nature Neuroscience* thanks Kenneth Harris and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

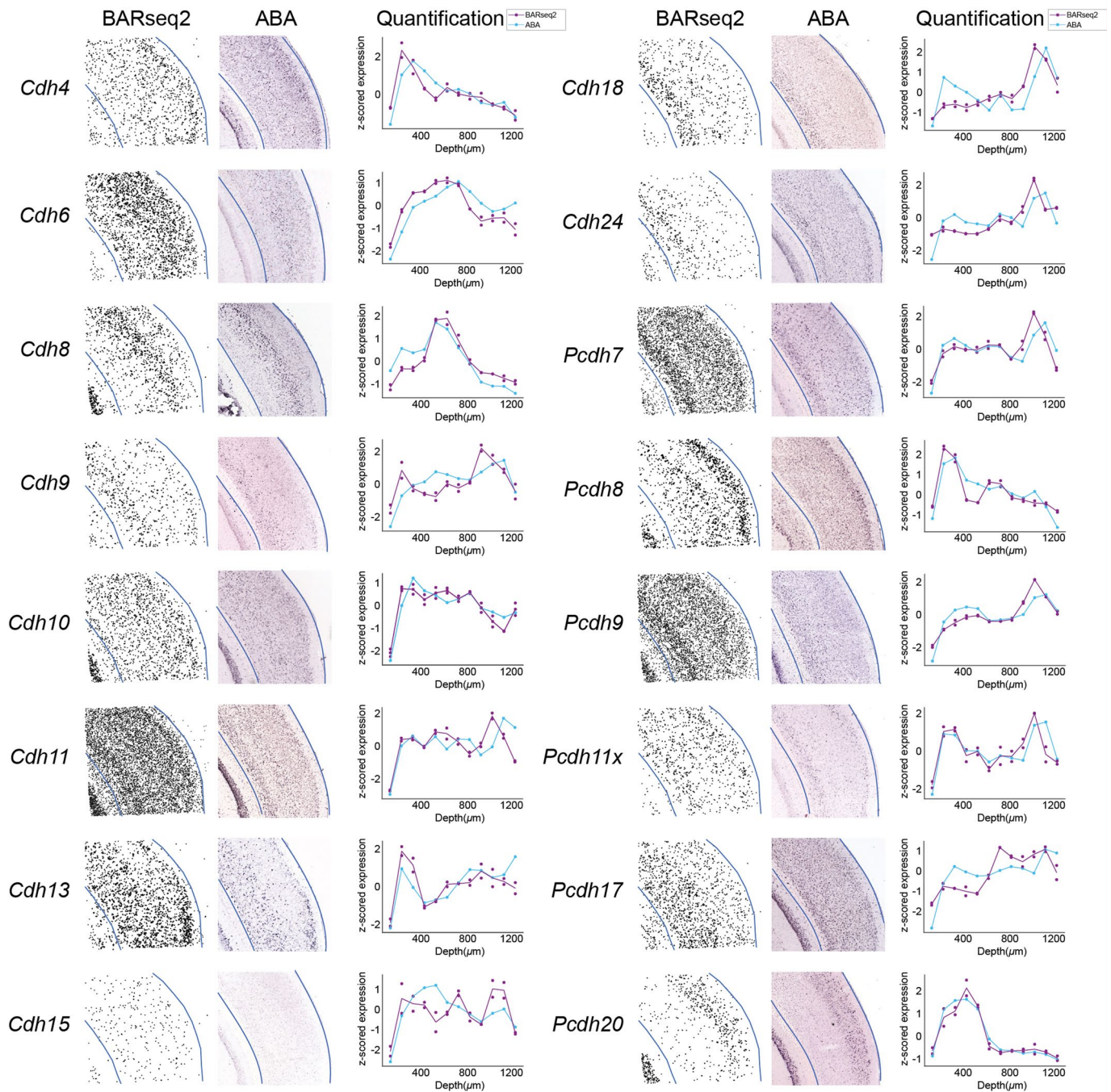
Reprints and permissions information is available at www.nature.com/reprints.



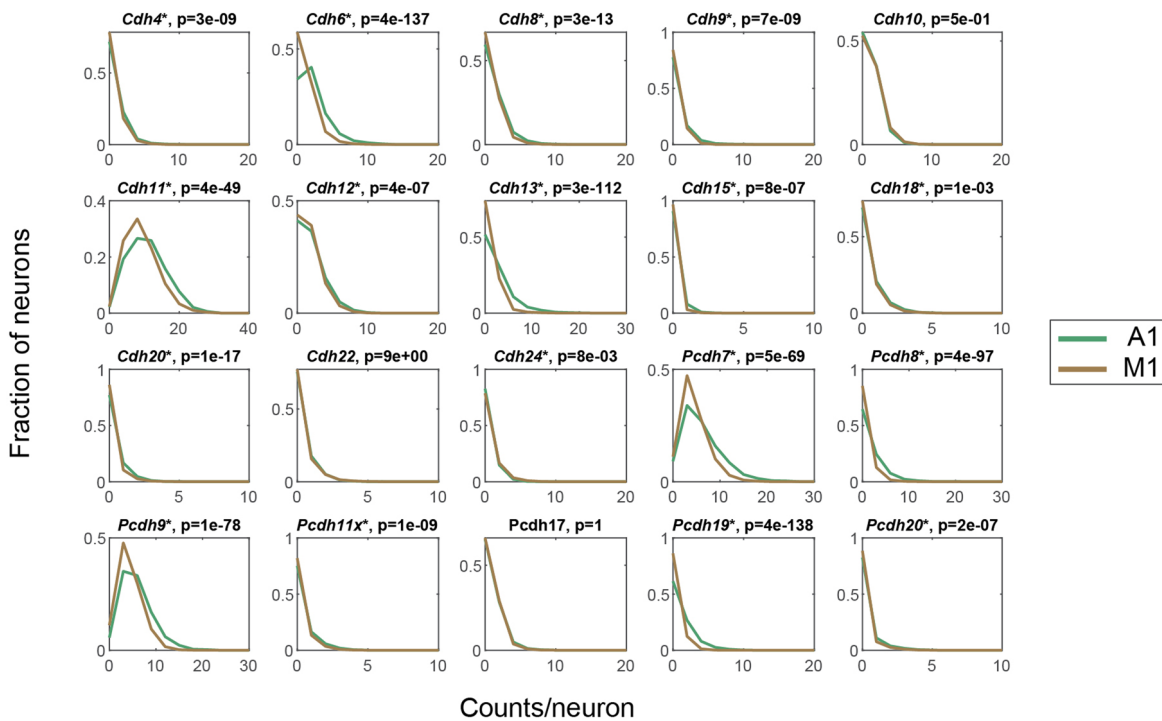
Extended Data Fig. 1 | Optimization of BARseq2 for detecting endogenous mRNAs. **a**, Relative sensitivity (means and individual data points) of BARseq2 in detecting *Slc17a7* using the indicated fixation times, normalized to that achieved with 5 mins of fixation. $n = 3$ for 480 mins and $n = 4$ for other conditions. **b**, Rolony counts for *Slc17a7* using either random primers or specific primers at two different concentrations. The two concentrations used were 5 μM (low) and 50 μM (high) for random primers, and 0.5 μM (low) and 5 μM (high) for specific primers. Lines indicate means and dots/crosses represent individual samples. $n = 2$ slices for each condition. **c,d**, BARseq2 sensitivity compared to RNAseq2. **c**, Spot density detected by BARseq2 or RNAseq2 in each 100 μm bin along the laminar axis in auditory cortex. Error bars indicate standard errors. The dashed line indicates linear fit for *Slc30a3* and *Cdh13*. Slope = 1.65 and $R^2 = 0.73$. $n = 5$ slices for both BARseq2 and RNAseq2. **d**, shows the means and individual samples for each gene. **e,f**, Positions of rolony across five sequencing cycles using the original (**e**) or the optimized (**f**) sequencing protocol. Scale bars = 10 μm. **g**, The distribution of minimum distance between rolony imaged in the first cycle and in the fifth cycle using the original or the optimized protocol. **h**, Median distance between rolony imaged in the indicated cycles and the closest rolony imaged in the first cycle using the original or the optimized protocol. Error bars indicate standard errors. For both (**g**) and (**h**), $n = 148,708$ rolony for optimized condition and $n = 12,114$ for original condition. **i,j**, The distribution of absolute rolony intensities for the first sequencing cycle (**i**) and relative rolony intensities after 6 sequencing cycles and one stripping step, normalized to the intensities in the first sequencing cycle (**j**). Amino-allyl dUTP concentrations used are indicated. In (**i**), $n = 63,852$ rolony for 0.08 μM and $n = 4,286$ rolony for 0.5 μM; in (**j**), $n = 128,976$ rolony for 0.08 μM and $n = 113,235$ rolony for 0.5 μM.



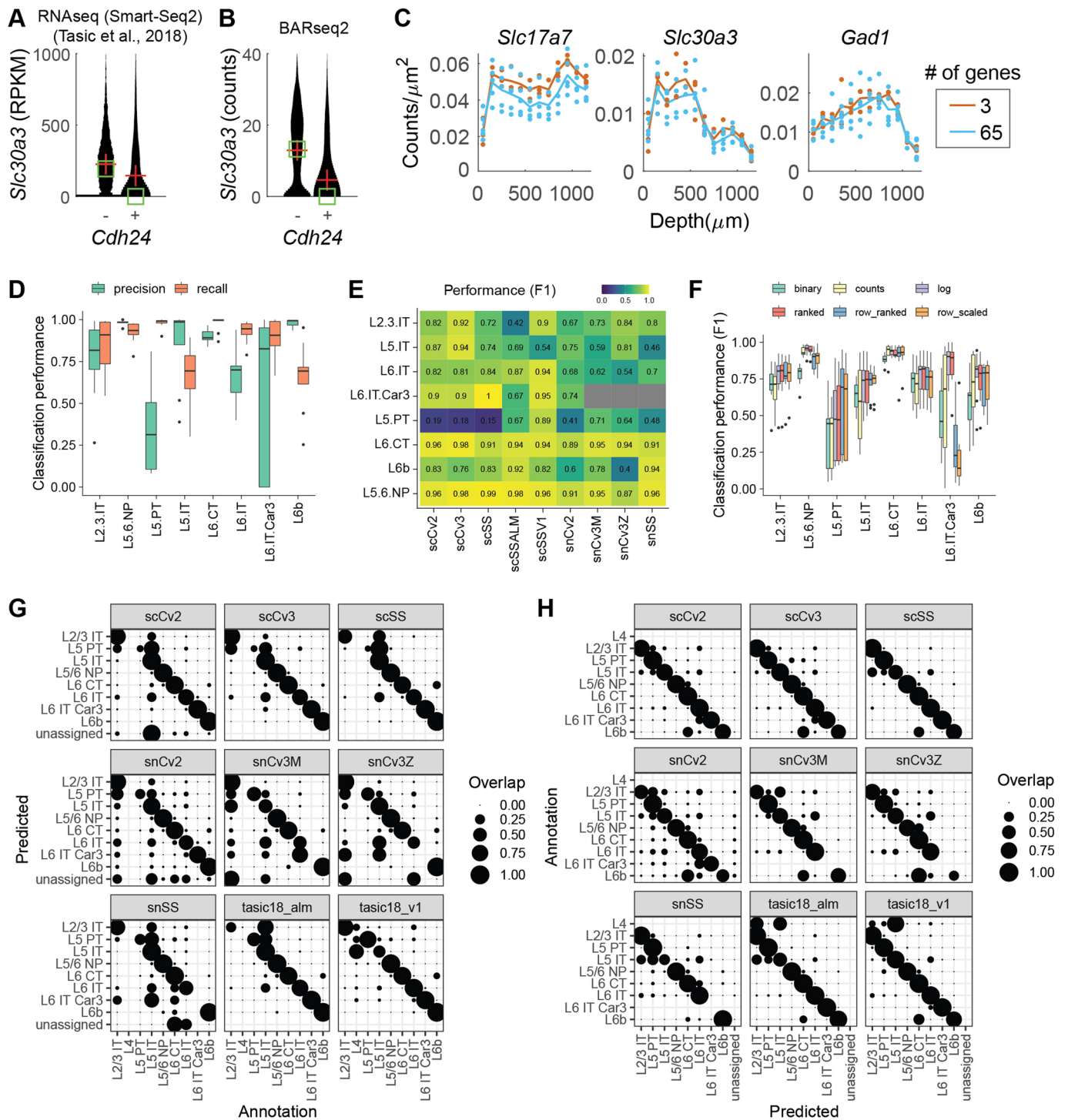
Extended Data Fig. 2 | Laminar distribution of cadherins in auditory cortex (green) and motor cortex (brown). In both cortical areas, cortical depth is normalized so that the bottom and the top of the cortex match between M1 and A1.



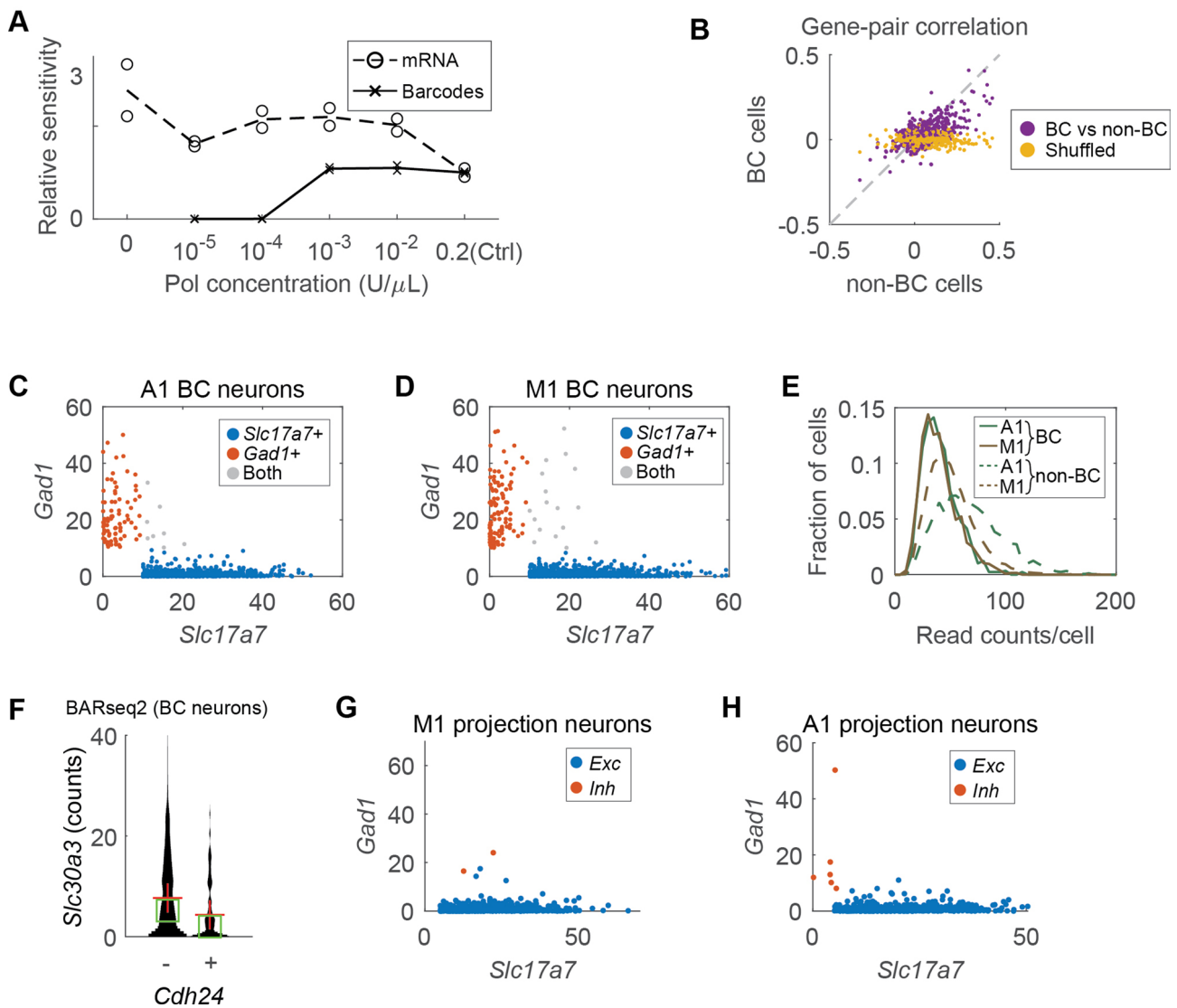
Extended Data Fig. 3 | Comparison between BARseq2 and Allen gene expression atlas. Gene expression patterns in auditory cortex identified by BARseq2 are plotted next to in situ hybridization images of the same genes in Allen gene expression atlas (ABA) and the quantified laminar distribution of the gene in both datasets. Only genes that had coronal images in the Allen gene expression atlas are shown. Blue lines indicate the boundaries of the cortex in both BARseq2 and ABA images. In the laminar distribution plots, dots represent values from two BARseq2 samples (purple) and one ABA sample (blue) per gene. Lines indicate means across samples.



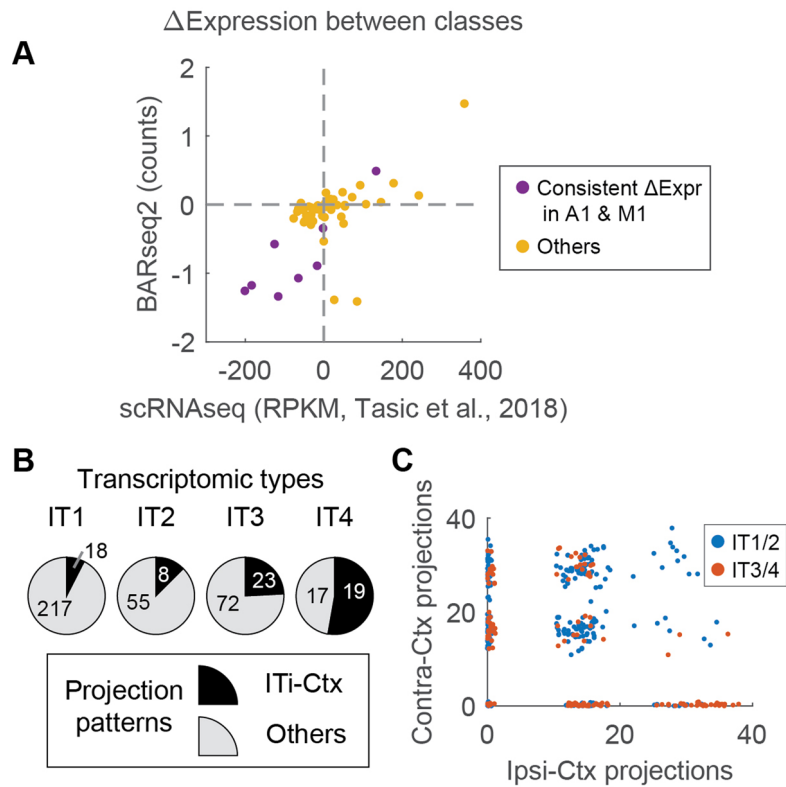
Extended Data Fig. 4 | The distribution of read counts per cell for the indicated genes in auditory cortex (green) and motor cortex (brown). Asterisks indicate genes with significant difference in expression between the two areas ($p < 0.05$ using two-tailed rank sum test after Bonferroni correction). p values after Bonferroni correction are indicated on top.



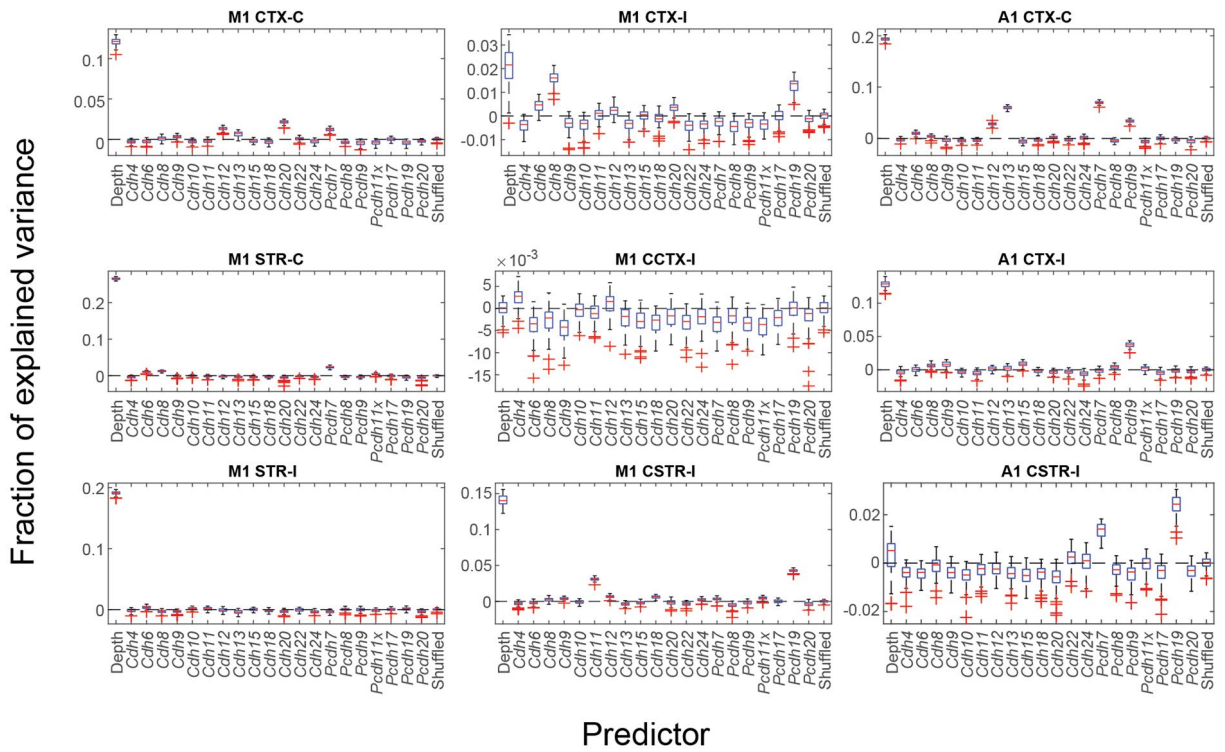
Extended Data Fig. 5 | Transcriptomic typing using BARseq2. **a, b**, *Slc30a3* expression in excitatory neurons with or without *Cdh24* expression in single-cell RNAseq (**a**) from Tasic, et al.³ or in BARseq2 (**b**). A cell is considered expressing *Cdh24* if the expression is higher than 10 RPKM in RNAseq or 1 count in BARseq2. Red crosses indicate means and green squares indicate medians. **c**, Expression density (means and individual data points) across laminar positions for the indicated genes. $n=3$ slices for the three-gene panel and $n=5$ slices for the 65-gene panel. **d**, Precision and recall of cell typing using the marker gene panel across nine single cell datasets. $N=9$ independent datasets shown in (**e**). In each box, the center shows the median, the bounds of the box show the 1st and 3rd quartiles, the whiskers show the range of the data, and points further than 1.5 IQR (Inter-Quartile Range) from the box are shown as outliers. **e**, Breakdown of average performance for each cell type in each dataset. The datasets are: scSSALM and scSSV1 are single cell SmartSeq datasets from ALM and V1 respectively³. All other datasets are BICCN M1 datasets²³ and the name indicates the technology used (sc = single cell, sn = single nuclei, Cv2/3 = Chromium v2/3, SS = SmartSeq). **f**, Average cell typing performance for six normalization strategies. $N=9$ independent datasets shown in (**e**). The box plots are generated in the same way as (**d**). **g**, Confusion matrix showing overlap between prediction and annotations, normalized by predictions. This plot emphasizes precision; it indicates the probability that a given prediction was correct. **h**, Confusion matrix showing overlap between prediction and annotations, normalized by annotations. This plot emphasizes recall; it indicates the probability that a given annotation was recovered.



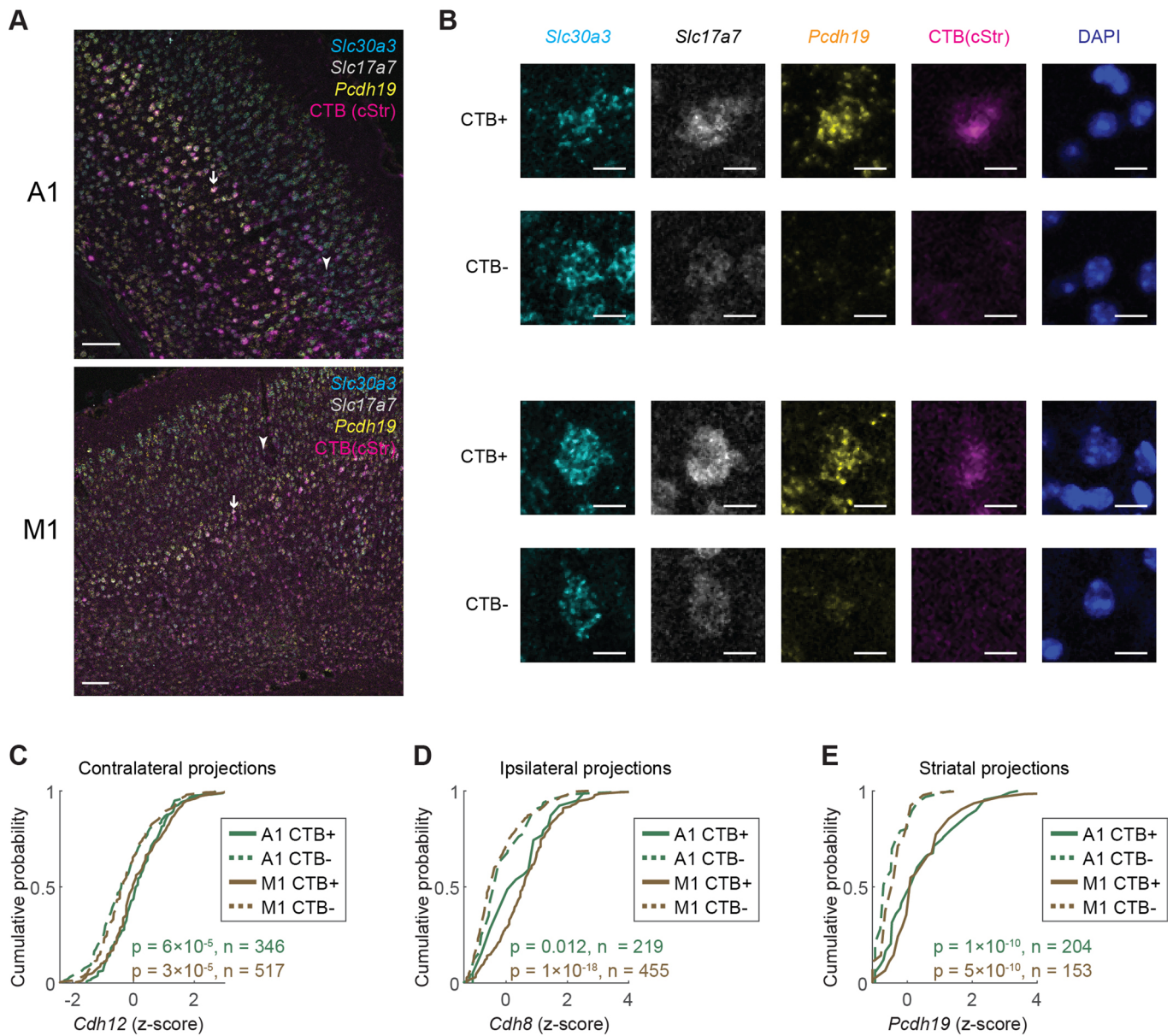
Extended Data Fig. 6 | Correlating gene expression to projections using BARseq2. **a**, Relative sensitivity of BARseq2 to barcodes (solid line) and endogenous mRNAs (dashed line) using the indicated concentration of Phusion DNA polymerase. Sensitivities are normalized to the original BARseq condition (*Ctrl*). Circles and crosses show individual data points across $n=2$ slices. **b**, Correlation between pairs of genes in barcoded cells (y-axis) and in non-barcoded cells (x-axis) as determined by BARseq2. Shuffled data (yellow) are also plotted for comparison. **c,d**, *Slc17a7* (x-axes) and *Gad1* (y-axes) expression in barcoded neurons in auditory (**c**) or motor cortex (**d**). Only neurons with more than 10 counts in either gene are shown. **e**, The distributions of read counts per barcoded neuron (solid lines) or non-barcoded neuron (dashed lines) in auditory (green) and motor (brown) cortex. **f**, *Slc30a3* expression in barcoded excitatory neurons with or without *Cdh24* expression in BARseq2. A cell is considered expressing *Cdh24* if the expression is higher than 1 count. Red crosses indicate means and green squares indicate median. **g,h**, *Slc17a7* (x-axes) and *Gad1* (y-axes) expression in barcoded projection neurons in motor (**g**) or auditory cortex (**h**). Excitatory and inhibitory neurons are color-coded as indicated.



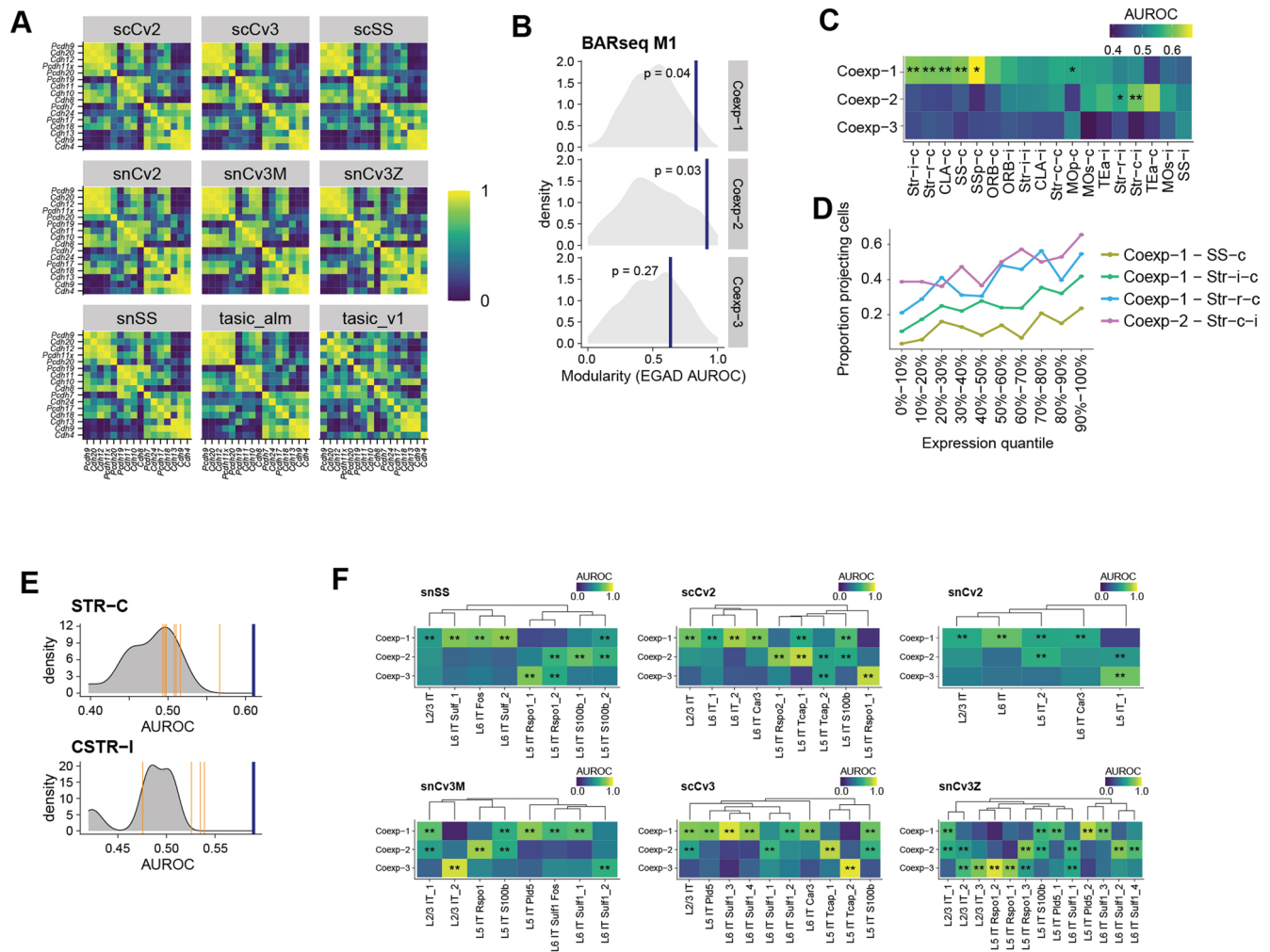
Extended Data Fig. 7 | BARseq2 reveals projection and gene expression differences across major classes and IT subtypes. **a**, Differential gene expression across major classes (IT, PT, and CT) observed using BARseq2 and single-cell RNAseq. Each dot shows the difference in mean expression of a gene across a pair of major classes observed using BARseq2 (y-axis) or single-cell RNAseq (x-axis). Differences in expression that were statistically significant (FDR < 0.05 using two-tailed rank sum tests) in both A1 and M1 as shown by BARseq2 are labeled purple; otherwise they are labeled yellow. The single-cell RNAseq data used were collected in the visual cortex and anterior-lateral motor cortex³. **b**, The fraction of ITi-Ctx neurons in four transcriptomic types of IT neurons in auditory cortex. ITi-Ctx neurons have only ipsilateral cortical projections and no striatal projections or contralateral projections⁶. The number of ITi-Ctx neurons and neurons with other projection patterns for each transcriptomic type are labeled on top of the pie charts. **c**, The projection strengths for contralateral (y-axis) and ipsilateral (x-axis) cortical projections for each IT neuron in auditory cortex. IT1/IT2 neurons are labeled blue and IT3/IT4 neurons are labeled red.



Extended Data Fig. 8 | Variance in projections explained by cadherins and laminar positions. Box plots of variance in each projection modules explained by the indicated predictors after 100 iterations of 10-fold cross validation. Boxes indicate second and third quartiles and whiskers indicate minimum and maximum values excluding outliers. Outliers are shown in red.



Extended Data Fig. 9 | Validation of correlation between cadherins and IT projections. **a**, Representative images of in situ hybridization in A1 (top) and M1 (bottom) slices with CTB labeling in the caudal striatum. Three marker genes and CTB labeling are shown in the indicated colors. Scale bars = 100 μ m. Arrows and arrowheads indicate example CTB+ and CTB- neurons, respectively. Experiments for each combination of targeted gene and CTB labeling condition (*Cdh12* with contralateral labeling, *Cdh8* with ipsilateral labeling, and *Pcdh19* with striatal labeling) were performed in slices from two animals. **b**, Crops of the indicated individual channels of example neurons from (a). Scale bars = 10 μ m. **c,d,e**, Cumulative probability distribution of the expression of *Cdh12* (c), *Cdh8* (d), and *Pcdh19* (e) in neurons with or without retrograde labeling of contralateral (c), ipsilateral (d), or caudal striatal (e) projections. p values from two-tailed rank sum tests after Bonferroni correction and numbers of neurons used for each experiment are indicated. N=2 animals for each experiment.



Extended Data Fig. 10 | Cadherin co-expression modules correlate with IT projections. **a**, Correlation among cadherins in IT neurons in motor cortex identified in the indicated single-cell RNAseq datasets^{3,23}. The datasets included are: tasic_alm and tasic_v1 are single cell SmartSeq datasets from ALM and V1 respectively³; all other datasets are BICCN M1 datasets²³; the name indicates the technology used (sc = single cell, sn = single nuclei, Cv2/3 = Chromium v2/3, SS = SmartSeq). **b**, Modularity (EGAD AUROC) of co-expression modules in BARseq2 M1 against null distribution of modularity (node permutation). BARseq2 modularity is shown by the blue lines with the corresponding p-values. P values are calculated using a one-sided non-parametric node permutation test without multiple comparison correction. **c**, Association (AUROC) between cadherin co-expression modules and the indicated projections. Significant associations are marked by asterisks (* FDR < 0.1, ** FDR < 0.05). **d**, Fractions of neurons with the indicated projections as a function of co-expression module expression. **e**, Distribution of associations of the indicated projection modules with gene expression. Association with significant gene module is shown by a blue line; association with single genes from that module is shown by orange lines; association with all other genes is shown by a gray density. **f**, Association of the three co-expression modules in transcriptomic IT neurons in the indicated datasets (AUROC, significance shown as in **c**).

Bibliography

- Adams, Mark D. et al. (Mar. 24, 2000). “The Genome Sequence of *Drosophila melanogaster*”. In: *Science* 287.5461. Publisher: American Association for the Advancement of Science, pp. 2185–2195. (Visited on 09/29/2021).
- Alon, Shahar et al. (Jan. 29, 2021). “Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems”. In: *Science* 371.6528. Publisher: American Association for the Advancement of Science Section: Research Article. (Visited on 06/23/2021).
- AlQuraishi, Mohammed (Dec. 8, 2020). *AlphaFold2 @ CASP14: “It feels like one’s child has left home.”* Some Thoughts on a Mysterious Universe. (Visited on 07/29/2021).
- Amezquita, Robert A. et al. (Feb. 2020). “Orchestrating single-cell analysis with Bioconductor”. In: *Nature Methods* 17.2. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 2 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Genomic analysis;Software Subject_term_id: genomic-analysis;software, pp. 137–145. (Visited on 06/27/2021).
- Andersson, Charlotte et al. (Nov. 1, 2019). “70-year legacy of the Framingham Heart Study”. In: *Nature Reviews Cardiology* 16.11. Publisher: Nature Publishing Group, pp. 687–698. (Visited on 01/22/2021).
- Arendt, Detlev et al. (2016). “The origin and evolution of cell types”. In: *Nature Publishing Group* 17. (Visited on 11/08/2017).
- Armstrong, Jacob et al. (2020). “Dynamic linkage of COVID-19 test results between Public Health England’s Second Generation Surveillance System and UK Biobank”. In: *Microbial Genomics* 6.7. Publisher: Microbiology Society, e000397. (Visited on 07/29/2021).

- Ashburner, Michael et al. (May 2000). “Gene Ontology: tool for the unification of biology”. In: *Nature genetics* 25.1, pp. 25–29. (Visited on 07/30/2021).
- Asp, Michaela, Joseph Bergenstråhle, and Joakim Lundeberg (May 4, 2020). “Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration”. In: *BioEssays*. Publisher: John Wiley and Sons Inc., p. 1900221. (Visited on 07/24/2020).
- Azimi, Nima et al. (June 1, 2017). “Discrepancies in stereotaxic coordinate publications and improving precision using an animal-specific atlas”. In: *Journal of Neuroscience Methods* 284. Publisher: Elsevier B.V., pp. 15–20. (Visited on 10/08/2020).
- Ballouz, Sara, Alexander Dobin, and Jesse A. Gillis (Dec. 2019). “Is it time to change the reference genome?” In: *Genome Biology* 20.1, p. 159. (Visited on 07/29/2021).
- Behdenna, Abdelkader et al. (Mar. 18, 2020). “pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods”. In: *bioRxiv*. Publisher: bioRxiv, p. 2020.03.17.995431. (Visited on 05/13/2021).
- Bendesky, Andres et al. (Apr. 27, 2017). “The genetic basis of parental care evolution in monogamous mice”. In: *Nature* 544.7651. Publisher: Nature Publishing Group, pp. 434–439. (Visited on 10/07/2020).
- Benoist, Christophe et al. (Oct. 1, 2012). “Consortium biology in immunology: The perspective from the Immunological Genome Project”. In: *Nature Reviews Immunology* 12.10. Publisher: Nature Publishing Group, pp. 734–740. (Visited on 02/21/2020).
- Bhattacharya, Jay and Mikko Packalen (Feb. 2020). *Stagnation and Scientific Incentives*. Cambridge, MA: National Bureau of Economic Research. (Visited on 02/24/2020).
- Biancalani, Tommaso et al. (Sept. 24, 2020). “Deep learning and alignment of spatially-resolved whole transcriptomes of single cells in the mouse brain with Tangram”. In: *bioRxiv*, p. 2020.08.29.272831. (Visited on 06/23/2021).
- Blibaum, Ash, Jonathan Werner, and Alexander Dobin (Nov. 10, 2019). “<p>STARsolo: single-cell RNA-seq analyses beyond gene expression</p>”. In: *F1000Research* 8. (Visited on 07/30/2021).
- Bohland, Jason W. et al. (Feb. 1, 2010). “Clustering of spatial gene expression patterns in the mouse brain and comparison with classical neuroanatomy”. In: *Methods* 50.2. Publisher: Academic Press, pp. 105–112. (Visited on 10/05/2018).

- Bol, Thijs, Mathijs de Vaan, and Arnout van de Rijt (May 8, 2018). “The Matthew effect in science funding”. In: *Proceedings of the National Academy of Sciences* 115.19. Publisher: National Academy of Sciences Section: Social Sciences, pp. 4887–4890. (Visited on 09/29/2021).
- Boulesteix, Anne Laure, Sabine Lauer, and Manuel J.A. Eugster (Apr. 24, 2013). “A Plea for Neutral Comparison Studies in Computational Sciences”. In: *PLoS ONE* 8.4. Publisher: Public Library of Science, p. 61562. (Visited on 06/04/2021).
- Bowman, S. et al. (May 29, 1997). “The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XIII”. In: *Nature* 387.6632, pp. 90–93.
- Bycroft, Clare et al. (Oct. 11, 2018). “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726. Publisher: Nature Publishing Group, pp. 203–209. (Visited on 01/25/2021).
- Cadwell, Cathryn R. et al. (Feb. 1, 2016). “Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq”. In: *Nature Biotechnology* 34.2. Publisher: Nature Publishing Group, pp. 199–203. (Visited on 04/17/2020).
- Canales, Roger D. et al. (Sept. 1, 2006). “Evaluation of DNA microarray results with quantitative gene expression platforms”. In: *Nature Biotechnology* 24.9. Publisher: Nature Publishing Group, pp. 1115–1122. (Visited on 10/07/2020).
- Cao, Junyue et al. (Aug. 18, 2017). “Comprehensive single-cell transcriptional profiling of a multicellular organism”. In: *Science* 357.6352. Publisher: American Association for the Advancement of Science Section: Research Article, pp. 661–667. (Visited on 08/18/2021).
- cap-example* (Nov. 1, 2020). original-date: 2019-11-21T21:59:19Z. (Visited on 07/29/2021).
- Chen, Ao et al. (Jan. 24, 2021a). “Large field of view-spatially resolved transcriptomics at nanoscale resolution Short title: DNA nanoball stereo-sequencing”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory. (Visited on 06/11/2021).
- Chen, Fei, Paul W. Tillberg, and Edward S. Boyden (Jan. 30, 2015). “Expansion microscopy”. In: *Science* 347.6221. Publisher: American Association for the Advancement of Science Section: Report, pp. 543–548. (Visited on 06/23/2021).

- Chen, Kok Hao et al. (2015). “Spatially resolved, highly multiplexed RNA profiling in single cells”. In: *Science*. (Visited on 11/16/2018).
- Chen, Shuonan et al. (Mar. 8, 2021b). “Barcode DEmixing through Non-negative Spatial Regression (BarDensr)”. In: *PLOS Computational Biology* 17.3. Publisher: Public Library of Science, e1008256. (Visited on 06/26/2021).
- Chen, Xiaoyin et al. (Nov. 28, 2017). “Efficient in situ barcode sequencing using padlock probe-based BaristaSeq”. In: *Nucleic Acids Research*. (Visited on 12/27/2017).
- Chen, Xiaoyin et al. (Oct. 17, 2019). “High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing”. In: *Cell* 179.3. Publisher: Cell Press, 772–786.e19. (Visited on 10/07/2020).
- Cho, Chun-seok et al. (2021). “Microscopic examination of spatial transcriptome using Seq-Scope”. In: *Cell*. Publisher: Elsevier Inc., pp. 1–14.
- Chon, Uree et al. (Dec. 1, 2019). “Enhanced and unified anatomical labeling for a common mouse brain atlas”. In: *Nature Communications* 10.1. Publisher: Nature Publishing Group, pp. 1–12. (Visited on 10/08/2020).
- Cleary, Brian et al. (Apr. 15, 2021). “Compressed sensing for highly efficient imaging transcriptomics”. In: *Nature Biotechnology*. Publisher: Nature Research, pp. 1–7. (Visited on 06/10/2021).
- Clevers, Hans et al. (2017). “What Is Your Conceptual Definition of Cell Type in the Context of a Mature Organism?” In: *Cell Systems* 4.3, pp. 255–259.
- Close, Jennie L., Brian R. Long, and Hongkui Zeng (Jan. 6, 2021). “Spatially resolved transcriptomics in neuroscience”. In: *Nature Methods* 18.1. Publisher: Nature Publishing Group, pp. 23–25. (Visited on 01/19/2021).
- Codeluppi, Simone et al. (Nov. 1, 2018). “Spatial organization of the somatosensory cortex revealed by osmFISH”. In: *Nature Methods* 15.11. Publisher: Nature Publishing Group, pp. 932–935. (Visited on 07/26/2020).
- Cohen, Ido et al. (2017). “DeepBrain: Functional Representation of Neural In-Situ Hybridization Images for Gene Ontology Classification Using Deep Convolutional Autoencoders”. In: Springer, Cham, pp. 287–296. (Visited on 10/05/2018).

- Crick, Francis and Edward Jones (1993). “Backwardness of human neuroanatomy”. In: *Nature* 361.6408, pp. 109–110.
- “Crystallography” (Oct. 1, 1971). “Crystallography: Protein Data Bank”. In: *Nature New Biology* 233.42. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 42 Primary_atype: News Publisher: Nature Publishing Group, pp. 223–223. (Visited on 09/28/2021).
- Deng, Yanxiang et al. (June 7, 2021a). “Spatial-ATAC-seq: spatially resolved chromatin accessibility profiling of tissues at genome scale and cellular level”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory. (Visited on 06/08/2021).
- Deng, Yanxiang et al. (Mar. 12, 2021b). “Spatial Epigenome Sequencing at Tissue Scale and Cellular Level”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2021.03.11.434985. (Visited on 06/21/2021).
- Di Bella, Daniela J. et al. (June 23, 2021). “Molecular logic of cellular diversification in the mouse cerebral cortex”. In: *Nature*, pp. 1–6. (Visited on 06/24/2021).
- Dobin, Alexander et al. (Jan. 2013). “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1, pp. 15–21. (Visited on 07/30/2021).
- Dries, Ruben et al. (Mar. 8, 2021). “Giotto: a toolbox for integrative analysis and visualization of spatial expression data”. In: *Genome Biology* 22.1, p. 78. (Visited on 06/26/2021).
- Economou, Michael N. et al. (Nov. 31, 2018). “Distinct descending motor cortex pathways and their roles in movement”. In: *Nature* 563.7729. Publisher: Nature Publishing Group, pp. 79–84. (Visited on 01/16/2019).
- Eddy, Sean R. (Apr. 8, 2013). “The ENCODE project: Missteps overshadowing a success”. In: *Current Biology* 23.7. Publisher: Elsevier, R259–R261. (Visited on 07/28/2021).
- Edgar, Ron, Michael Domrachev, and Alex E. Lash (Jan. 1, 2002). “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository”. In: *Nucleic Acids Research* 30.1, pp. 207–210. (Visited on 07/29/2021).
- Edsgård, Daniel, Per Johnsson, and Rickard Sandberg (Mar. 19, 2018). “Identification of spatial expression trends in single-cell gene expression data”. In: *Nature Methods* 15.5. Publisher: Nature Publishing Group, pp. 339–342. (Visited on 05/02/2018).

- Eisen, Michael B. et al. (1998). “Cluster analysis and display of genome-wide expression patterns”. In: *Proceedings of the National Academy of Sciences* 95.25, pp. 14863–14868.
- Eiss, Robert (Aug. 25, 2020). “Confusion over Europe’s data-protection law is stalling scientific progress”. In: *Nature* 584.7822. Bandiera_abtest: a Cg_type: World View Number: 7822 Publisher: Nature Publishing Group Subject_term: Law, Policy, Research data, pp. 498–498. (Visited on 07/30/2021).
- Ellrott, Kyle et al. (Sept. 10, 2019). “Reproducible biomedical benchmarking in the cloud: lessons from crowd-sourced data challenges”. In: *Genome Biology* 20.1, p. 195. (Visited on 09/29/2021).
- Emmert-Buck, M. R. et al. (Nov. 8, 1996). “Laser capture microdissection”. In: *Science (New York, N.Y.)* 274.5289, pp. 998–1001.
- Eng, Chee-Huat Linus et al. (Apr. 2019). “Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+”. In: *Nature* 568.7751, pp. 235–239. (Visited on 06/22/2021).
- Evert, Stefan et al. (2016). “Outliers or Key Profiles? Understanding Distance Measures for Authorship Attribution”. In: *Conference Abstracts. Digital Humanities 2016*. Jagiellonian University & Pedagogical University, Kraków, pp. 188–191.
- Fazal, Furqan M. et al. (July 11, 2019). “Atlas of Subcellular RNA Localization Revealed by APEX-Seq”. In: *Cell* 178.2. Publisher: Cell Press, 473–490.e26. (Visited on 07/24/2020).
- Fischer, Stephan and Jesse Gillis (June 30, 2021). “How many markers are needed to robustly determine a cell’s type?” In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2021.04.16.439807. (Visited on 07/01/2021).
- Fornito, Alex, Aurina Arnatkevi, and Ben D Fulcher (2018). “Bridging the Gap between Connectome and Transcriptome”. In: *Trends in Cognitive Sciences* xx. (Visited on 11/26/2018).
- French, Leon and Paul Pavlidis (Jan. 6, 2011). “Relationships between Gene Expression and Brain Wiring in the Adult Rodent Brain”. In: *PLoS Computational Biology*

- 7.1. Ed. by Olaf Sporns. Publisher: Public Library of Science, e1001049. (Visited on 10/07/2018).
- French, Leon, Powell Patrick Cheng Tan, and Paul Pavlidis (July 29, 2011). “Large-Scale Analysis of Gene Expression and Connectivity in the Rodent Brain: Insights through Data Integration”. In: *Frontiers in Neuroinformatics* 5. Publisher: Frontiers, p. 12. (Visited on 10/07/2018).
- Friedman, Lauren G. et al. (Jan. 1, 2015). “Cadherin-8 expression, synaptic localization and molecular control of neuronal form in prefrontal cortico-striatal circuits”. In: *The Journal of comparative neurology* 523.1, pp. 75–92. (Visited on 07/16/2021).
- Fu, Xiaonan et al. (Mar. 17, 2021). “Continuous Polony Gels for Tissue Mapping with High Resolution and RNA Capture Efficiency”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 2021.03.17.435795. (Visited on 06/11/2021).
- Fulcher, Ben D and Alex Fornito (Feb. 2, 2016). “A transcriptional signature of hub connectivity in the mouse connectome.” In: *Proceedings of the National Academy of Sciences of the United States of America* 113.5. Publisher: National Academy of Sciences, pp. 1435–40. (Visited on 10/07/2018).
- Fuzik, János et al. (Feb. 2016). “Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes”. In: *Nature Biotechnology* 34.2, pp. 175–183. (Visited on 07/11/2021).
- Fürth, Daniel, Victor Hatini, and Je H. Lee (Aug. 5, 2019). “In Situ Transcriptome Accessibility Sequencing (INSTA-seq)”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 722819. (Visited on 06/22/2021).
- Fürth, Daniel et al. (Jan. 4, 2018). “An interactive framework for whole-brain maps at cellular resolution”. In: *Nature Neuroscience* 21.1. Publisher: Nature Publishing Group, pp. 139–149. (Visited on 02/06/2018).
- Ganguli, Deep, Ambrose Carr, and Brian Long (Dec. 17, 2018). *Developing a Computational Pipeline and Benchmark Datasets for Image-Based Transcriptomics*. ASCB. (Visited on 06/26/2021).

- GenomeAsia100K Consortium (Dec. 5, 2019). “The GenomeAsia 100K Project enables genetic discoveries across Asia”. In: *Nature* 576.7785, pp. 106–111. (Visited on 07/29/2021).
- Giacomello, Stefania et al. (May 8, 2017). “Spatially resolved transcriptome profiling in model plant species”. In: *Nature Plants* 3.6, pp. 1–11. (Visited on 06/18/2021).
- Gouwens, Nathan W. et al. (Nov. 12, 2020). “Integrated Morphoelectric and Transcriptomic Classification of Cortical GABAergic Cells”. In: *Cell* 183.4. Publisher: Elsevier, 935–953.e19. (Visited on 07/11/2021).
- Grange, Pascal, Michael Hawrylycz, and Partha P. Mitra (Mar. 14, 2013). “Computational neuroanatomy and co-expression of genes in the adult mouse brain, analysis tools for the Allen Brain Atlas”. In: *Quantitative Biology* 1.1. Publisher: Springer-Verlag, pp. 91–100. (Visited on 10/05/2018).
- Grange, Pascal and Partha P. Mitra (2012). “Computational neuroanatomy and gene expression: Optimal sets of marker genes for brain regions”. In: *2012 46th Annual Conference on Information Sciences and Systems, CISS 2012*. ISBN: 978-1-4673-3140-1. (Visited on 05/23/2021).
- Grange, Pascal et al. (Apr. 8, 2014). “Cell-type-based model explaining coexpression patterns of genes in the brain.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.14. Publisher: National Academy of Sciences, pp. 5397–402. (Visited on 10/07/2018).
- Hall, Mark A (1999). “Correlation-based Feature Selection for Machine Learning”. PhD thesis. The University of Waikato. (Visited on 08/01/2018).
- Halpern, Keren Bahar et al. (Sept. 17, 2018). “Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells”. In: *Nature Biotechnology*. Publisher: Nature Publishing Group. (Visited on 09/18/2018).
- Hanchate, Naresh K. et al. (Feb. 25, 2020). “Connect-seq to superimpose molecular on anatomical neural circuit maps”. In: *Proceedings of the National Academy of Sciences of the United States of America* 117.8. Publisher: National Academy of Sciences, pp. 4375–4384. (Visited on 10/07/2020).

- Hanley, James A. and Barbara J. McNeil (1982). “The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve”. In: *Radiology*, pp. 29–36.
- Hawrylycz, Mike et al. (Nov. 1, 2011). “Multi-scale correlation structure of gene expression in the brain”. In: *Neural Networks* 24.9. Publisher: Pergamon, pp. 933–942. (Visited on 10/07/2018).
- Hayano, Yuki et al. (Dec. 15, 2014). “The role of T-cadherin in axonal pathway formation in neocortical circuits”. In: *Development* 141.24, pp. 4784–4793. (Visited on 07/16/2021).
- Heath, Allison P. et al. (Mar. 2021). “The NCI Genomic Data Commons”. In: *Nature Genetics* 53.3, pp. 257–262. (Visited on 07/29/2021).
- Henry, Alex M. and John G. Hohmann (Oct. 26, 2012). “High-resolution gene expression atlases for adult and developing mouse brain and spinal cord”. In: *Mammalian Genome* 23.9. Publisher: Springer-Verlag, pp. 539–549. (Visited on 10/07/2018).
- Hintiryan, Hourii et al. (Aug. 1, 2016). “The mouse cortico-striatal projectome”. In: *Nature Neuroscience* 19.8. Publisher: Nature Publishing Group, pp. 1100–1114. (Visited on 05/02/2021).
- Hitti, Frederick L. and Steven A. Siegelbaum (2014). “The hippocampal CA2 region is essential for social memory”. In: *Nature* 508.1. Publisher: Nature Publishing Group, pp. 88–92. (Visited on 01/19/2021).
- Horstmann, D. M. (1991). “The Sabin live poliovirus vaccination trials in the USSR, 1959.” In: *The Yale Journal of Biology and Medicine* 64.5, pp. 499–512. (Visited on 07/30/2021).
- Hu, Jian et al. (Dec. 2, 2020a). “Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.11.30.405118. (Visited on 06/23/2021).
- Hu, Kenneth H. et al. (July 6, 2020b). “ZipSeq: barcoding for real-time mapping of single cell transcriptomes”. In: *Nature Methods*. Publisher: Nature Research, pp. 1–11. (Visited on 07/26/2020).

- Huang, Longwen et al. (July 9, 2020). “BRICseq Bridges Brain-wide Interregional Connectivity to Neural Activity and Gene Expression in Single Animals”. In: *Cell* 182.1. Publisher: Cell Press, 177–188.e27. (Visited on 10/07/2020).
- Huang, Z. Josh (2014). “Toward a genetic dissection of cortical circuits in the mouse”. In: *Neuron* 83.6. Publisher: Cell Press, pp. 1284–1302. (Visited on 05/23/2021).
- Johnson, W. Evan, Cheng Li, and Ariel Rabinovic (Jan. 1, 2007). “Adjusting batch effects in microarray expression data using empirical Bayes methods”. In: *Biostatistics* 8.1. Publisher: Oxford Academic, pp. 118–127. (Visited on 05/13/2021).
- Junker, Jan Philipp et al. (Oct. 23, 2014). “Genome-wide RNA Tomography in the Zebrafish Embryo”. In: *Cell* 159.3. Publisher: Cell Press, pp. 662–675. (Visited on 07/26/2020).
- Karaiskos, Nikos et al. (Oct. 13, 2017). “The Drosophila embryo at single-cell transcriptome resolution”. In: *Science* 358.6360. Publisher: American Association for the Advancement of Science, pp. 194–199. (Visited on 07/25/2020).
- Ke, Rongqin et al. (Sept. 2013). “In situ sequencing for RNA analysis in preserved tissue and cells”. In: *Nature Methods* 10.9. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cancer genetics;Gene expression;Gene expression analysis;Sequencing Subject_term_id: cancer-genetics;gene-expression;gene-expression-analysis;sequencing, pp. 857–860. (Visited on 06/22/2021).
- Kebschull, Justus M. et al. (Sept. 7, 2016). “High-Throughput Mapping of Single-Neuron Projections by Sequencing of Barcoded RNA”. In: *Neuron* 91.5. Publisher: Cell Press, pp. 975–987. (Visited on 12/26/2017).
- Kleshchevnikov, Vitalii et al. (Nov. 17, 2020). *Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics*. preprint. Genomics. (Visited on 06/24/2021).
- Ko, Younhee et al. (Feb. 19, 2013). “Cell type-specific genes show striking and distinct patterns of spatial expression in the mouse brain.” In: *Proceedings of the National Academy of Sciences of the United States of America* 110.8. Publisher: National Academy of Sciences, pp. 3095–100. (Visited on 10/07/2018).

- Kodama, Yuichi, Martin Shumway, and Rasko Leinonen (Jan. 2012). “The sequence read archive: explosive growth of sequencing data”. In: *Nucleic Acids Research* 40 (Database issue), pp. D54–D56. (Visited on 07/29/2021).
- Krzanowski, Wojtek J. and David J. Hand (2009). *ROC Curves for Continuous Data*. CRC Press Taylor & Francis Group.
- Laeremans, Annelies et al. (Jan. 2013). “AMIGO2 mRNA expression in hippocampal CA2 and CA3a”. In: *Brain Structure and Function* 218.1. Publisher: Brain Struct Funct, pp. 123–130. (Visited on 05/02/2021).
- Langmead, Ben et al. (Mar. 4, 2009). “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. In: *Genome Biology* 10.3, R25. (Visited on 07/25/2021).
- Lee, Je Hyuk et al. (2015). “Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues”. In: *Nature Protocols*. ISBN: 1750-2799 (Electronic)\r1750-2799 (Linking).
- Leek, Jeffrey T. et al. (Mar. 2012). “The SVA package for removing batch effects and other unwanted variation in high-throughput experiments”. In: *Bioinformatics* 28.6. Publisher: Oxford University Press, pp. 882–883. (Visited on 05/13/2021).
- Lein, Ed, Lars E Borm, and Sten Linnarsson (Oct. 6, 2017). “The promise of spatial transcriptomics for neuroscience in the era of molecular cell typing.” In: *Science (New York, N.Y.)* 358.6359. Publisher: American Association for the Advancement of Science, pp. 64–69. (Visited on 09/08/2018).
- Lein, Ed S., Xinyu Zhao, and Fred H. Gage (Apr. 14, 2004). “Defining a Molecular Atlas of the Hippocampus Using DNA Microarrays and High-Throughput In Situ Hybridization”. In: *Journal of Neuroscience* 24.15. Publisher: Society for Neuroscience, pp. 3879–3889. (Visited on 05/02/2021).
- Lein, Ed S. et al. (Jan. 6, 2007). “Genome-wide atlas of gene expression in the adult mouse brain”. In: *Nature* 445.7124. Publisher: Nature Publishing Group, pp. 168–176. (Visited on 09/28/2018).
- Lein, Edward S. et al. (Apr. 25, 2005). “Redefining the boundaries of the hippocampal CA2 subfield in the mouse using gene expression and 3-dimensional reconstruction”.

- In: *Journal of Comparative Neurology* 485.1. Publisher: J Comp Neurol, pp. 1–10. (Visited on 05/02/2021).
- Levsky, Jeffrey M. and Robert H. Singer (July 15, 2003). “Fluorescence in situ hybridization: past, present and future”. In: *Journal of Cell Science* 116.14, pp. 2833–2838. (Visited on 10/01/2021).
- Liu, Yang et al. (Dec. 10, 2020). “High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue”. In: *Cell* 183.6. Publisher: Cell Press, 1665–1681.e18. (Visited on 06/08/2021).
- Longo, Sophia K. et al. (June 18, 2021). “Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics”. In: *Nature Reviews Genetics*, pp. 1–18. (Visited on 06/30/2021).
- Lovatt, Ditte et al. (Feb. 12, 2014). “Transcriptome in vivo analysis (TIVA) of spatially defined single cells in live tissue”. In: *Nature Methods* 11.2. Publisher: Nature Publishing Group, pp. 190–196. (Visited on 10/19/2018).
- Lu, Shaina et al. (July 19, 2021). “Assessing the replicability of spatial gene expression using atlas data from the adult mouse brain”. In: *PLOS Biology* 19.7. Publisher: Public Library of Science, e3001341. (Visited on 08/02/2021).
- Lubeck, Eric et al. (Apr. 2014). “Single cell in situ RNA profiling by sequential hybridization”. In: *Nature methods* 11.4, pp. 360–361. (Visited on 06/22/2021).
- Lundmark, Anna et al. (2018). “Gene expression profiling of periodontitis-affected gingival tissue by spatial transcriptomics”. In: (visited on 05/28/2019).
- MacKenzie-Graham, Allan et al. (Feb. 2004). “A multimodal, multidimensional atlas of the C57BL/6J mouse brain”. In: *Journal of Anatomy* 204.2. Publisher: Wiley-Blackwell, pp. 93–102. (Visited on 10/07/2020).
- Mahmood, Syed S. et al. (Mar. 15, 2014). “The Framingham Heart Study and the epidemiology of cardiovascular disease: A historical perspective”. In: *The Lancet* 383.9921. Publisher: Lancet Publishing Group, pp. 999–1008. (Visited on 01/22/2021).
- Marx, Vivien (Jan. 6, 2021). “Method of the Year: spatially resolved transcriptomics”. In: *Nature Methods* 18.1. Publisher: Nature Publishing Group, pp. 9–14. (Visited on 01/19/2021).

- Mason, S J and N E Graham (2002). “Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves”. In: *Quarterly Journal of the Royal Meteorological Society* 128, pp. 2145–2166.
- Matsunaga, Eiji et al. (2015). “Complex and dynamic expression of cadherins in the embryonic marmoset cerebral cortex”. In: *Development, Growth & Differentiation* 57.6. [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/dgd.12228](https://onlinelibrary.wiley.com/doi/pdf/10.1111/dgd.12228), pp. 474–483. (Visited on 07/16/2021).
- Maynard, Kristen R et al. (June 19, 2020). “dotdotdot: an automated approach to quantify multiplex single molecule fluorescent in situ hybridization (smFISH) images in complex tissues”. In: *Nucleic Acids Research* 48.11, e66–e66. (Visited on 06/26/2021).
- Maynard, Kristen R. et al. (Mar. 2021). “Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex”. In: *Nature Neuroscience* 24.3. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 3 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Gene expression;Molecular neuroscience Subject_term_id: gene-expression;molecular-neuroscience, pp. 425–436. (Visited on 08/18/2021).
- Medaglia, Chiara et al. (Dec. 22, 2017). “Spatial reconstruction of immune niches by combining photoactivatable reporters and scRNA-seq”. In: *Science* 358.6370. Publisher: American Association for the Advancement of Science, pp. 1622–1626. (Visited on 07/26/2020).
- Merritt, Christopher R. et al. (May 1, 2020). “Multiplex digital spatial profiling of proteins and RNA in fixed tissue”. In: *Nature Biotechnology* 38.5. Publisher: Nature Research, pp. 586–599. (Visited on 07/24/2020).
- Merton, Robert K. (Jan. 5, 1968). “The Matthew Effect in Science”. In: *Science* 159.3810. Publisher: American Association for the Advancement of Science, pp. 56–63. (Visited on 09/29/2021).
- “Meta-analysis in basic biology” (Dec. 1, 2016). In: *Nature Methods* 13.12. Publisher: Nature Publishing Group, p. 959. (Visited on 10/08/2020).

- Moffitt, Jeffrey R. et al. (Nov. 16, 2018). “Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region”. In: *Science* 362.6416. Publisher: American Association for the Advancement of Science. (Visited on 10/07/2020).
- Molnár-Gábor, Fruzsina and Jan O Korbel (Mar. 6, 2020). “Genomic data sharing in Europe is stumbling—Could a code of conduct prevent its fall?” In: *EMBO Molecular Medicine* 12.3, e11421. (Visited on 07/30/2021).
- Moor, Andreas E et al. (2018). “Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis”. In: *Cell* 175. (Visited on 09/28/2018).
- Moses, Lambda and Lior Pachter (May 12, 2021). *Museum of Spatial Transcriptomics*. preprint. Bioinformatics. (Visited on 06/25/2021).
- Navarro, José Fernández et al. (Aug. 15, 2017). “ST Pipeline: an automated pipeline for spatial mapping of unique transcripts”. In: *Bioinformatics* 33.16, pp. 2591–2593. (Visited on 06/25/2021).
- Network (BICCN), BRAIN Initiative Cell Census et al. (Oct. 21, 2020). “A multimodal cell census and atlas of the mammalian primary motor cortex”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.10.19.343129. (Visited on 07/22/2021).
- Ng, Lydia et al. (2007). “Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4.3. ISBN: 1545-5963, pp. 382–392.
- Niedzwiedz, Claire L. et al. (May 29, 2020). “Ethnic and socioeconomic differences in SARS-CoV-2 infection: prospective cohort study using UK Biobank”. In: *BMC Medicine* 18.1, p. 160. (Visited on 07/29/2021).
- Nitzan, Mor et al. (Dec. 2019). “Gene expression cartography”. In: *Nature* 576.7785, pp. 132–137.
- Ortiz, Cantin et al. (June 1, 2020). “Molecular atlas of the adult mouse brain”. In: *Science Advances* 6.26. Publisher: American Association for the Advancement of Science, eabb3446. (Visited on 10/07/2020).

- Padrón, Alejandro, Shintaro Iwasaki, and Nicholas T. Ingolia (Aug. 22, 2019). “Proximity RNA Labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules”. In: *Molecular Cell* 75.4. Publisher: Cell Press, 875–887.e5. (Visited on 07/24/2020).
- Palla, Giovanni et al. (Apr. 29, 2021). “Squidpy: a scalable framework for spatial single cell analysis”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2021.02.19.431994. (Visited on 06/25/2021).
- Paul, Anirban et al. (Oct. 19, 2017). “Transcriptional Architecture of Synaptic Communication Delineates GABAergic Neuron Identity”. In: *Cell* 171.3, 522–539.e20.
- Pedregosa, Fabian et al. (2011). *Scikit-learn: Machine Learning in Python*. Publication Title: Journal of Machine Learning Research Volume: 12 Issue: 85 ISSN: 1533-7928, pp. 2825–2830. (Visited on 10/08/2020).
- Perkel, Jeffrey M. (Aug. 19, 2019). “Starfish enterprise: finding RNA patterns in single cells”. In: *Nature* 572.7770. Bandiera_abtest: a Cg_type: Technology Feature Number: 7770 Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics, Transcriptomics, Imaging, Technology, pp. 549–551. (Visited on 06/26/2021).
- Pham, Duy et al. (May 31, 2020). “stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.05.31.125658. (Visited on 07/04/2021).
- Phillips, James W. et al. (Nov. 2019). “A repeated molecular architecture across thalamic pathways”. In: *Nature Neuroscience* 22.11. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 11 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Genetics of the nervous system;Molecular neuroscience;Neural circuits;Neuroscience Subject_term_id: genetics-of-the-nervous-system;molecular-neuroscience;neural-circuit;neuroscience, pp. 1925–1935. (Visited on 07/06/2021).

- Poulin, Jean Francois et al. (Sept. 1, 2016). “Disentangling neural cell diversity using single-cell transcriptomics”. In: *Nature Neuroscience* 19.9. Publisher: Nature Publishing Group, pp. 1131–1141. (Visited on 10/08/2020).
- Press Secretary, Office of the (June 26, 2000). *June 2000 White House Event*. Genome.gov. (Visited on 07/29/2021).
- Purves, Dale et al. (2001a). “Functional Organization of the Primary Motor Cortex”. In: *Neuroscience. 2nd edition*. Publisher: Sinauer Associates. (Visited on 10/01/2021).
- (2001b). “The Auditory Cortex”. In: *Neuroscience. 2nd edition*. Publisher: Sinauer Associates. (Visited on 10/01/2021).
- Qiu, Shenfeng et al. (2012). “Single-neuron RNA-Seq: technical feasibility and reproducibility”. In: *Frontiers in Genetics* 3. Publisher: Frontiers. (Visited on 07/11/2021).
- Regev, Aviv et al. (Dec. 5, 2017). “The Human Cell Atlas”. In: *eLife* 6. Ed. by Thomas R Gingeras. Publisher: eLife Sciences Publications, Ltd, e27041. (Visited on 07/04/2021).
- Regev, Aviv et al. (Oct. 11, 2018). “The Human Cell Atlas White Paper”. In: (visited on 01/16/2019).
- Righelli, Dario et al. (Jan. 27, 2021). “SpatialExperiment: infrastructure for spatially resolved transcriptomics data in R using Bioconductor”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2021.01.27.428431. (Visited on 06/27/2021).
- Roberts, Kenny et al. (Mar. 20, 2021). “Transcriptome-wide spatial RNA profiling maps the cellular architecture of the developing human neocortex”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 2021.03.20.436265. (Visited on 06/10/2021).
- Robinson, Mark D. and Olga Vitek (Oct. 9, 2019). “Benchmarking comes of age”. In: *Genome Biology* 20.1. Publisher: BioMed Central Ltd., p. 205. (Visited on 10/08/2020).
- Rodrigues, Samuel G. et al. (Mar. 29, 2019). “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”. In: *Science* 363.6434. Publisher: American Association for the Advancement of Science, pp. 1463–1467. (Visited on 04/08/2019).

- Satija, Rahul et al. (May 13, 2015). “Spatial reconstruction of single-cell gene expression data”. In: *Nature Biotechnology* 33.5. Publisher: Nature Publishing Group, pp. 495–502. (Visited on 01/31/2019).
- Scala, Federico et al. (2020). “Phenotypic variation within and across transcriptomic cell types in mouse motor cortex”. In: (visited on 04/17/2020).
- Sever, Richard et al. (Nov. 6, 2019). “bioRxiv: the preprint server for biology”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 833400. (Visited on 07/29/2021).
- Shah, Sheel et al. (Oct. 19, 2016). “In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus”. In: *Neuron* 92.2. Publisher: Elsevier, pp. 342–357. (Visited on 06/22/2021).
- Shi, Leming et al. (Sept. 1, 2006). “The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements”. In: *Nature Biotechnology* 24.9. Publisher: Nature Publishing Group, pp. 1151–1161. (Visited on 10/07/2020).
- Singer, R. H. and D. C. Ward (Dec. 1, 1982). “Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog”. In: *Proceedings of the National Academy of Sciences* 79.23, pp. 7331–7335.
- Skininder, Michael A., Jordan W. Squair, and Leonard J. Foster (May 2019). “Evaluating measures of association for single-cell transcriptomics”. In: *Nature Methods* 16.5, pp. 381–386. (Visited on 07/13/2021).
- Slippery Mystery | Radiolab* (May 27, 2020). WNYC Studios. (Visited on 07/29/2021).
- Srivatsan, Sanjay R. et al. (July 2, 2021). “Embryo-scale, single-cell spatial transcriptomics”. In: *Science* 373.6550, pp. 111–117. (Visited on 08/17/2021).
- Stickels, Robert R. et al. (Dec. 7, 2020). “Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2”. In: *Nature Biotechnology*. Publisher: Nature Research, pp. 1–7. (Visited on 01/19/2021).
- Stock, M et al. (Dec. 1, 2017). “Maintaining and disseminating the kilogram following its redefinition”. In: *Metrologia* 54.6, S99–S107. (Visited on 07/30/2021).

- Stuart, Tim et al. (June 13, 2019). “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177.7, 1888–1902.e21. (Visited on 06/24/2021).
- Ståhl, Patrik L et al. (July 1, 2016). “Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.” In: *Science (New York, N.Y.)* 353.6294. Publisher: American Association for the Advancement of Science, pp. 78–82. (Visited on 12/29/2017).
- Su, Zhenqiang et al. (Sept. 2014). “A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium”. In: *Nature Biotechnology* 32.9. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 9 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Microarray analysis;RNA sequencing Subject_term_id: microarray-analysis;rna-sequencing, pp. 903–914. (Visited on 10/01/2021).
- Sun, Shiquan, Jiaqiang Zhu, and Xiang Zhou (Feb. 2020). “Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies”. In: *Nature Methods* 17.2. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 2 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Gene expression profiling;Software;Statistical methods;Transcriptomics Subject_term_id: gene-expression-profiling;software;statistical-methods;transcriptomics, pp. 193–200. (Visited on 06/28/2021).
- Sun, Yu-Chi et al. (May 10, 2021). “Integrating barcoded neuroanatomy with spatial transcriptional profiling enables identification of gene correlates of projections”. In: *Nature Neuroscience*. Publisher: Springer Science and Business Media LLC, pp. 1–13. (Visited on 05/27/2021).
- Svensson, Valentine, Sarah A. Teichmann, and Oliver Stegle (Apr. 27, 2018). “SpatialDE: Identification of spatially variable genes”. In: *Nature Methods* 15.5. Publisher: Nature Publishing Group, pp. 343–346. (Visited on 07/26/2020).
- Takei, Yodai et al. (Apr. 27, 2021a). *Integrated spatial genomics in tissues reveals invariant and cell type dependent nuclear architecture*. preprint. Systems Biology. (Visited on 06/22/2021).

- Takei, Yodai et al. (Feb. 11, 2021b). “Integrated spatial genomics reveals global architecture of single nuclei”. In: *Nature* 590.7845. Publisher: Nature Research, pp. 344–350. (Visited on 06/10/2021).
- Tan, Powell Patrick Cheng, Leon French, and Paul Pavlidis (Feb. 4, 2013). “Neuron-Enriched Gene Expression Patterns are Regionally Anti-Correlated with Oligodendrocyte-Enriched Patterns in the Adult Mouse and Human Brain”. In: *Frontiers in Neuroscience* 7. Publisher: Frontiers, p. 5. (Visited on 10/07/2018).
- Tasic, Bosiljka et al. (Nov. 31, 2018). “Shared and distinct transcriptomic cell types across neocortical areas”. In: *Nature* 563.7729. Publisher: Nature Publishing Group, pp. 72–78. (Visited on 01/16/2019).
- The C. Elegans Sequencing Consortium (Dec. 11, 1998). “Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology”. In: *Science* 282.5396. Publisher: American Association for the Advancement of Science, pp. 2012–2018. (Visited on 09/29/2021).
- The Gene Ontology Consortium et al. (Jan. 8, 2021). “The Gene Ontology resource: enriching a GOld mine”. In: *Nucleic Acids Research* 49 (D1), pp. D325–D334. (Visited on 07/30/2021).
- The International Brain Laboratory et al. (May 20, 2021). “Standardized and reproducible measurement of decision-making in mice”. In: *eLife* 10. Ed. by Naoshige Uchida and Michael J Frank. Publisher: eLife Sciences Publications, Ltd, e63711. (Visited on 07/28/2021).
- The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2008*. NobelPrize.org. (Visited on 10/02/2021).
- Tibshirani, Robert (1996). “Regression shrinkage and selection via the LASSO”. In: *Journal of the Royal Statistical Society B* 58.1, pp. 267–288.
- Toda, Chitoku et al. (Feb. 10, 2017). “POMC Neurons: From Birth to Death”. In: *Annual Review of Physiology* 79. Publisher: Annual Reviews Inc., pp. 209–236. (Visited on 01/19/2021).

- Todd, Annabel E. et al. (May 20, 2005). “Progress of Structural Genomics Initiatives: An Analysis of Solved Target Structures”. In: *Journal of Molecular Biology* 348.5, pp. 1235–1260. (Visited on 09/29/2021).
- Toga, Arthur W. et al. (Dec. 2006). “Towards multimodal atlases of the human brain”. In: *Nature Reviews Neuroscience* 7.12. Publisher: Nature Publishing Group, pp. 952–966. (Visited on 05/27/2021).
- Vickovic, Sanja et al. (Oct. 1, 2019). “High-definition spatial transcriptomics for in situ tissue profiling”. In: *Nature Methods* 16.10. Publisher: Nature Publishing Group, pp. 987–990. (Visited on 10/07/2020).
- Vickovic, Sanja et al. (Oct. 15, 2020). “SM-Omics: An automated platform for high-throughput spatial multi-omics”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.10.14.338418. (Visited on 06/23/2021).
- Vliet, Simon van et al. (Apr. 25, 2018). “Spatially Correlated Gene Expression in Bacterial Groups: The Role of Lineage History, Spatial Gradients, and Cell-Cell Interactions”. In: *Cell Systems* 6.4. Publisher: Elsevier, 496–507.e6. (Visited on 06/18/2021).
- Walt, Stéfan van der et al. (June 19, 2014). “scikit-image: image processing in Python”. In: *PeerJ* 2. Publisher: PeerJ Inc., e453. (Visited on 06/27/2021).
- Wang, Fay et al. (Jan. 2012). “RNAscope”. In: *The Journal of Molecular Diagnostics : JMD* 14.1, pp. 22–29. (Visited on 06/22/2021).
- Wang, Guiping, Jeffrey R. Moffitt, and Xiaowei Zhuang (Mar. 19, 2018). “Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy”. In: *Scientific Reports* 8.1, p. 4847. (Visited on 06/22/2021).
- Wang, Quanxin et al. (May 14, 2020). “The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas”. In: *Cell* 181.4. Publisher: Cell Press, 936–953.e20. (Visited on 10/08/2020).
- Wang, Xiao et al. (July 27, 2018). “Three-dimensional intact-tissue sequencing of single-cell transcriptional states”. In: *Science* 361.6400. Publisher: American Association for the Advancement of Science Section: Research Article. (Visited on 06/22/2021).

- Weinberg, Alvin M (1961). *Impact of Large-Scale Science on the United States*. Publication Title: Science Volume: 134 Issue: 3473 ISBN: 202016:38:37, pp. 161–164. (Visited on 02/21/2020).
- Weinstein, Joshua A., Aviv Regev, and Feng Zhang (June 27, 2019). “DNA Microscopy: Optics-free Spatio-genetic Imaging by a Stand-Alone Chemical Reaction”. In: *Cell* 178.1. Publisher: Cell Press, 229–241.e16. (Visited on 07/25/2020).
- Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis (Feb. 6, 2018). “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome Biology* 19.1, p. 15. (Visited on 06/27/2021).
- Wolf, Lior et al. (May 5, 2011). “Gene Expression in the Rodent Brain is Associated with Its Regional Connectivity”. In: *PLoS Computational Biology* 7.5. Ed. by Olaf Sporns. Publisher: Public Library of Science, e1002040. (Visited on 10/15/2018).
- Xia, Chenglong et al. (Sept. 24, 2019). “Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.39. Publisher: National Academy of Sciences, pp. 19490–19499. (Visited on 02/20/2020).
- Yang, Xiaoyu et al. (2020). “SMART-Q: An Integrative Pipeline Quantifying Cell Type-Specific RNA Transcription”. In: *PloS One* 15.4, e0228760.
- Yao, Zizhen et al. (Mar. 5, 2020). “An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types”. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 2020.02.29.970558. (Visited on 10/08/2020).
- Yao, Zizhen et al. (June 10, 2021). “A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation”. In: *Cell* 184.12, 3222–3241.e26. (Visited on 07/23/2021).
- Yates, Thomas et al. (Oct. 1, 2020). “Obesity and risk of COVID-19: analysis of UK biobank”. In: *Primary Care Diabetes* 14.5. Publisher: Elsevier, pp. 566–567. (Visited on 07/29/2021).

-
- Zeisel, Amit et al. (2018). “Molecular Architecture of the Mouse Nervous System Resource Molecular Architecture of the Mouse Nervous System”. In: *Cell* 174, pp. 999–1014. (Visited on 08/09/2018).
- Zhang, Zhuzhu et al. (2020). “Epigenomic Diversity of Cortical Projection Neurons in the Mouse Brain”. In: *bioRxiv*. (Visited on 04/16/2020).